Test Time Adaptation Methods for Point Cloud Registration in Laparoscopic Surgery

Nina Bodelot, Soufiane Belharbi, and Eric Granger

LIVIA, Dept. of Systems Engineering, ETS Montreal, Canada
nina.bodelot.1@ens.etsmtl.ca, eric.granger@etsmtl.ca

Abstract. 3D point cloud registration in laparoscopic surgery estimates the transformation between an intraoperative organ reconstructed from video and its preoperative mesh. Since ground-truth transformations are unavailable for real data, supervised networks are trained on synthetic organ pairs. At test time, however, real reconstructions differ from synthetic ones and are noisy, sparse and occluded, degrading correspondence estimation. Test-time adaptation (TTA) methods can address this domain shift at inference and have been applied successfully on tasks such as classification and segmentation. However, these methods typically rely on logits, entropy, class prototypes, or cache memory mechanisms unavailable in registration. Moreover, TTA methods assume a single shifted input, whereas registration involves a pair with an asymmetric shift, which mainly affects the intraoperative cloud, making TTA for registration much more challenging. This paper provides state-of-the-art TTA methods for 3D registration tasks across three families: model, normalization, and input adaptation. In particular, we analyze and modify four representative methods from those families that perform: auxiliary-task model update, backpropagation-free token purging, feature alignment, and layer-normalization calibration. TTA methods are modified to account for the asymmetric domain shifts between the preoperative and intraoperative point clouds, and to replace the entropy-based mechanism from classification. Given a correspondence-based model trained on clean source synthetic data, two scenarios were considered for the target: synthetic data with corruption and real data. We experiment on the P2P and P2ILReg datasets. Eight corruptions were applied separately to those datasets on the synthetic target data only, such as uniform noise and global density decrease, with five increasing levels of severity. Our experiments show that all methods improve registration on the P2P dataset, whereas on P2ILReg only input adaptation reduces the error, while normalization adaptation degrades it. Given the computational overhead of the backpropagation-based method, input adaptation is a more promising family for laparoscopic surgery, with low inference latency and a consistent reduction in error across all datasets.
Our code: https://github.com/ninaa-git/survey_pc_registration_tta

Keywords: Test-Time Adaptation · Point Cloud Registration · Laparoscopic Surgery · Corrupted Synthetic and Real Data.

1 Introduction

Laparoscopic registration enhances surgical visualization by aligning a preoperative 3D organ mesh with a partial intraoperative surface reconstructed from 2D video frames [14,19,42]. Supervised models establish correspondences to estimate rigid transformations [14,37,42] but require ground-truth labels unavailable during surgery. Source models are therefore trained on synthetic datasets and evaluated on real target data [21]. However, synthetic and real point clouds can differ in point density, visible surface, organ scale, and noise, introducing a domain shift between source and target. This shift can lead to incorrect registration, misleading the surgeon. Domain adaptation methods aim to mitigate such simulated-to-real distribution shift. Source-free domain adaptation (SFDA) uses unlabeled target data without source access [6,11,24]. Despite its success, it remains impractical for laparoscopy as surgical data is unavailable for adaptation. In contrast, TTA adapts the model during inference from the incoming target data [18,33], making it more suitable for surgery. We consider three families: model adaptation updates model parameters by backpropagation [9], normalization adaptation adapts normalization statistics [28], and input adaptation modifies test embeddings before prediction [5]. We study four representative methods of these families: Point-TTA [12] for model adaptation, Layer Normalization (LN) [39] for normalization adaptation, Progressive Embedding Alignment (PEA) [20] and Purge-Gate [38] for input adaptation. Despite its suitability, TTA remains limited for registration. Most methods were designed for classification [30,31] or segmentation [23], relying on logits, entropy, class prototypes, or cache memory, limiting their transfer for rigid transformation tasks. Applying these TTA methods to registration therefore requires changes. First, classification TTA adapts one target input. In registration, the target is a pair in which the shift mainly affects the intraoperative cloud. A single-input method estimates one correction and applies it to the whole pair, introducing a shift in the preoperative cloud. We therefore modify each method to account for this asymmetry. Then, logit-based criterion from classification do not transfer to registration tasks that output a transformation. It is therefore replaced.

Key contributions: (1) An analysis and modification of four representative methods across three TTA families (Point-TTA, LN, PEA, Purge-Gate) for 3D registration. They are modified to account for the asymmetric domain shift between the preoperative and intraoperative point clouds. We also replace the classification entropy-based mechanism of Purge-Gate with an inlier ratio mechanism. **(2)** A comparative evaluation of four TTA methods for laparoscopic point cloud registration on P2P and P2ILReg datasets [35,42], with Point-TTA on P2P only. Experiments are performed under two target-domain scenarios (corrupted synthetic and real intraoperative data), under episodic and continual settings. Our results show that the adaptation methods consistently minimize registration error on the P2P dataset, whereas input adaptation reduces the error, and normalization adaptation degrades performance on the P2ILReg dataset. Model adaptation introduces high computational overhead. Input adaptation family improves performance across all datasets and has a low inference latency, making it more suitable for laparoscopy surgery.

2 Related Work

(a) Point Cloud Registration: Deep rigid registration models extract and match embeddings to estimate global transformations [41]. Among these models, the supervised frameworks are broadly split into one-stage and two-stage architectures. One-stage methods directly estimate the transformation from the input points, whereas two-stage methods first match superpoints before computing the global transformation [34]. For laparoscopy surgery, both one-stage registration networks and superpoint-based methods have been proposed [36,42]. Among supervised methods, PARE-Net [37] is an effective approach [41], which is two-stage. It extracts point embeddings with specific convolutions and uses transformers to improve point-cloud matching [37].
(b) Test-Time Adaptation for 3D Point Cloud Registration: Test-time adaptation methods have mainly been explored in 2D [28], and they underperform when directly applied to 3D settings [25]. While 3D methods have mostly focused on classification [30,31], only one method specifically addresses point cloud registration [12]. Existing TTA approaches can be classified into four families according to the adapted component: the weights, normalization layers, input or predictions [18,33]. They can also be episodic [12,18], where each batch is adapted independently, or online [20,33], where the adaptation from previous target batches is reused for the following batches.

The first family, model adaptation, relies on auxiliary tasks or on test-time objectives. Auxiliary-task methods update an encoder shared between the main task and a self-supervised task, such as rotation prediction [9], contrastive learning [1], or reconstruction [12,16,22]. Objective-based methods directly update the model's weights at test time by minimizing prediction entropy [28], enforcing feature alignment [17], or optimizing against pseudo-labels [26]. Then, the normalization family updates only layer or batch normalization statistics to realign the target data with the source domain knowledge embedded in these layers [28,2,3,39]. Although lightweight, this strategy requires accurate target distribution estimation that has been best achieved in continual settings with strategies such as weighting [13] or class-balanced cache [10]. The input adaptation family aligns target inputs or embeddings with the source domain [5,25]. This improves source model performance as the input or embeddings resemble the source data distribution. While it avoids any model updates and can be applied to all model architectures, it can fail under severe domain shifts if the input adaptation is insufficient. Inference adaptation modifies the model output, for example through caching using logits [26] or combining different model predictions [29,40]. While this method improves the output predictions, it leaves the input and model untouched, which fails under drastic data shifts.

However, for all families, most of these methods are designed for classification and segmentation tasks [18,26]. The direct application of those families to rigid

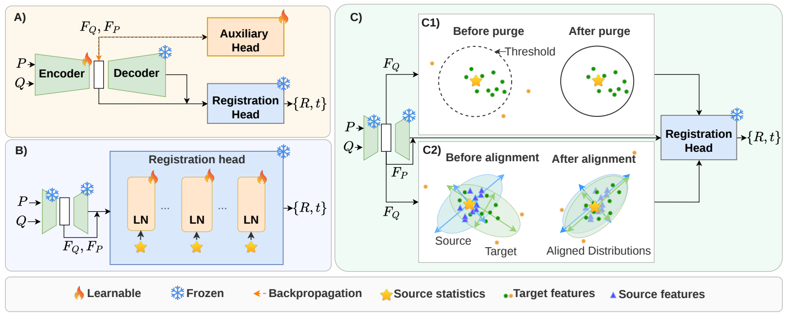

Fig. 1. Modified TTA methods, with P, Q: preoperative and intraoperative point clouds, $\mathcal{T} = \{R, t\}$: transformation, F_P, F_Q: output of the encoder. (A) Model adaptation using Point-TTA [12] relies on an auxiliary head. (B) Normalization adaptation modifies the normalization mechanism, for LN [39] using source and target statistics. (C) Input adaptation before the registration head: (C1) Purge-Gate [38] removes the divergent target features, (C2) PEA [20] aligns target features to source statistics.

registration remains limited. It has been explored in model adaptation with 3D auxiliary tasks [12] on indoor and outdoor scenes.

3 TTA Methods for Registration

This section describes our modifications to representative methods for registration (Fig. 1). In laparoscopic surgery, the shift mainly affects the intraoperative point cloud, whereas the preoperative point cloud is processed like the source mesh and is far less affected.

A) Point-TTA [12] is a model adaptation method allowing the update of weights at test time. The registration model contains a shared feature encoder, a primary registration branch, and an auxiliary branch. The shared encoder is adapted with backpropagation using self-supervised auxiliary losses. These losses are defined from auxiliary tasks such as point cloud reconstruction, feature learning, and correspondence classification [12]. Point-TTA uses a joint main and auxiliary task training for initialization, followed by a meta-auxiliary learning strategy [7] so that the auxiliary update benefits the primary registration task. At test time, the auxiliary loss updates the shared backbone for each sample, while the registration head remains frozen since no target registration labels are available. This enables sample-specific adaptation of the feature extractor. However, the source training is a joint stage followed by a meta training stage before deployment. Its main limitation for time-constrained registration is its high computational cost needed for several forward and backward passes.

Modifications: The reconstruction auxiliary task is used since it is the most effective auxiliary task in registration [12]. We perform a two-stage point cloud registration and use the encoder embeddings of the coarse point matching for coarse point cloud reconstruction. Since the encoder is shared between the intraoperative and preoperative clouds, adaptation uses the embeddings of both intraoperative and preoperative point clouds rather than only those of the intraoperative cloud.

B) Layer Normalization [39]: Normalization calibration methods restrict adaptation parameters of normalization layers, as they contain source scaling and shifting statistics. In LN [39], the pre-affine normalization is preserved, while the post-affine output is adapted using source and target moments. The target moments can be estimated using one batch only in an episodic setting, which can be unreliable, or over a window of batches. Although fast, this method limits the adaptation capacity.

Modifications: LN [39] was designed for classification, where each LN layer processes a single input and stores one source statistic. In registration, both the preoperative and intraoperative clouds pass through the same LN layers, in self-attention within each cloud and in cross-attention between them, but the shift mainly affects the intraoperative one. A statistic shared across the two clouds would therefore re-shift the preoperative cloud toward the intraoperative one. We instead split each adapted LN layer's statistics by cloud, storing separate source and target moments for each, and calibrate every cloud with its own. The target moments are estimated online.

C) Input Adaptation: This family adapts the input data before processing it while leaving the model unchanged ((C) in Fig.1). For instance, Purge-Gate [38] (C1 in Fig.1) removes tokens whose embeddings are the most divergent from a source-domain prototype. Removing outlier tokens allows the model predictions to be more accurate. The method is lightweight and episodic. However, the purging ratio is selected by minimizing classification entropy. This criterion is linked to classification logits and must therefore be modified for registration. Another work within this family is PEA [20](C2 in Fig.1). Instead of removing tokens, it aligns target embeddings toward source feature statistics. Source mean and covariance are computed offline, while target statistics are obtained online and used for covariance alignment. The aligned embeddings are then passed to the frozen registration head. PEA is designed to update its target statistics in a continual mode that relies on an exponential moving average (EMA) with momentum. PEA is backpropagation-free and does not modify the model weights. However, the computation of covariance matrices and their decomposition can introduce additional cost, especially when applied at several layers.

Modifications: Adaptation is applied to the embeddings of the intraoperative point cloud only, as it is the point cloud that contains the main shift. For Purge-Gate[38], we replace the entropy-based selection with the Inlier Ratio (IR), a common unsupervised measure of correspondence quality in registration [34,41]. We select the purging ratio $\mathcal{L}_{pg}$ that maximizes $\mathrm{IR}(\hat{T},\hat{\mathcal{C}}) = \frac{1}{k}\sum_{i=1}^{k}\mathbb{1}(\|\hat{R}\mathbf{p}_i + \hat{t} - \mathbf{q}_i\|_2 < \tau)$, where $(\mathbf{p}_i, \mathbf{q}_i)$ are matched correspondences and τ is the acceptance radius. For PEA, as for Purge-Gate, we only adapt the embeddings associated with the intraoperative point cloud. This differs from the original PEA, where the encoder is adapted progressively. In our case, adapting the encoder is not possible as the alignment would also shift the embeddings of the preoperative point cloud, which is far less affected by the shift. For PEA

[20] and LN [39], we consider statistics estimated either with an episodic (EP) strategy, where statistics are estimated per sample or a continual strategy (CT) where statistics are accumulated across samples. Both methods are designed for continual adaptation as statistical estimation with one sample is unreliable. LN averages statistics over a window [39], and PEA uses EMA of step m [20]. For a fair comparison, we use a modified EMA for both LN and PEA. In the EMA formulation, the first batch estimation is biased when it contains one sample. Let n be the first cumulative number of embeddings. We compute the target statistics as a cumulated mean [32], and switch to EMA once its update falls below the momentum m.

4 Results and Discussion

4.1 Experimental Methodology

Datasets and corruptions: Two synthetic datasets are used for source model training, and three target sets for testing (two corrupted synthetic, one real). The first synthetic dataset, P2P [35], contains preoperative livers and intraoperative liver that are cropped to visible surfaces of the preoperative liver. We split the original training set into 80%/20% for training and validation for 1048/261 samples, and test on the provided 6315 liver pairs. The second one, P2ILReg [42], provides synthetic and real test sets with preoperative meshes. The real test set has 92 intraoperative point clouds reconstructed from individual laparoscopic video key frames. The synthetic set is a 60%/20%/20% train/validation/test split, provided by [42] and totaling 29400/9800/9800 samples. To emulate target-domain shifts, corruptions are applied only to the synthetic intraoperative target point clouds. The real test set remains unchanged, since real data inherently introduces a domain shift. We use noise corruptions: uniform (uni), Gaussian (gauss), background (backg), impulse (impul); and sampling/visibility corruptions: global density decrease (g-d-dec), local density decrease (l-d-dec), cutout (cut), occlusion (oclsion). Each corruption is generated at five levels of increasing severity, following point cloud corruption benchmarks [27] and laparoscopic liver registration settings [42], with results averaged over severity levels.
Evaluation metrics: For synthetic data, we report relative rotation error (RRE) and relative translation error (RTE), following [42]. We additionally report root mean squared error (RMSE) directly, rather than the registration recall used in [42] to avoid thresholding RMSE at a specific distance. For real intraoperative data, ground-truth rigid transformations are not available. Following [42], we report Chamfer Distance (CD)[4] between the point clouds, and the Dice score between the projected registered organ surface and the annotated organ mask.
Implementation details: PARE-Net [37] is used as the correspondence-based registration backbone. Following the dataset protocols [35,42], training is performed for 150/80 epochs on P2P/P2ILReg. Batch size is 12 for source training and 1 for testing, as key frames are processed one at a time. Experiments are run on NVIDIA H100 (P2P) and NVIDIA RTX A6000 (P2ILReg). For Point-TTA, the learning rate for the auxiliary task is $\gamma = 1.6 \times 10^{-4}$, and $K = 5$ steps to

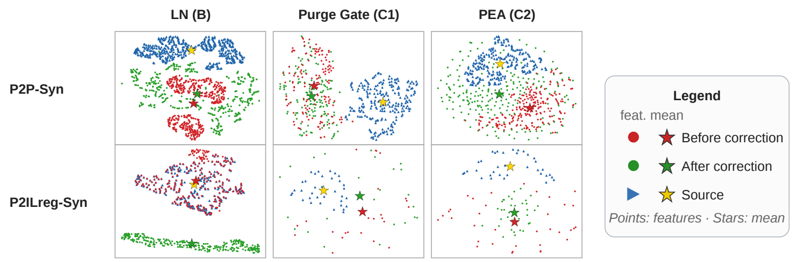

Fig. 2. t-SNE for uniform noise 5. For LN and PEA, embeddings appear before (red) and after (green) correction. For Purge-Gate, tokens are purged (red) or kept (green).

update the backbone as in [12]. Point-TTA is only evaluated on the synthetic P2P data [35] as its joint and meta training on P2ILReg is computationally prohibitive. For modified methods, TTA hyperparameters are selected per dataset by minimizing RMSE on the source validation set rather than the corrupted test data. This avoids test label leakage, but may be suboptimal since the validation data is clean. For Purge-Gate [38], the purging ratio is selected among $\mathcal{L}_{pg} \in \{0.0, 0.1, 0.5\}$. For continual mode, PEA uses $m = 0.05$ for P2P dataset [35] and $m = 0.02$ for P2ILReg [42], and LN uses $m = 0.01$ on both datasets.

4.2 Comparison of State-of-the-Art Methods

Table 1 shows that all modified methods improve registration on P2P-Syn, led by LN (-5.41 episodic, -5.50 continual). On the more challenging P2ILReg-Syn dataset, whose source baseline is higher (38.70 mm vs 15.47 mm), LN degrades registration (+8.61). Indeed, under large shifts, LN target statistics become unreliable and move features away from the source. PEA and Purge-Gate remain robust (-1.80, -1.63). PEA registration is degraded notably for background noise on both datasets. Purge-Gate improves registration on both datasets and for any corruption, as it selects the purge ratio by inlier ratio and keeps the unpurged input when purging would lower it. Table 2 confirms with RRE/RTE that PEA and Purge-Gate improve registration across the synthetic datasets. However, the gains remain small, around 2° and 1mm, which is limited for surgery relative to the source-only errors of 21°/15mm on P2P-Syn and 36°/24mm on P2ILReg-Syn. On P2ILReg-Real, LN again degrades registration. Purge-Gate and continual PEA improve registration on CD/Dice, from 4.26mm/71.87 to 4.17mm/72.96 and 4.22mm/72.22, respectively. PEA episodic is close to the source and improves Dice to 72.86. As with the synthetic data, improvements remain limited. Unlike LN, Purge-Gate and continual PEA improve registration over all datasets. On synthetic datasets, the continual mode improves LN while improving (P2ILReg-Syn) or degrading (P2P-Syn) PEA. On the real dataset, it degrades LN; and on PEA improves CD by 0.05 mm and degrades the Dice score by 0.64. Samples on synthetic and real datasets come from unordered samples or non-consecutive frames, respectively, from different patients. Accumulated target statistics may therefore be too disparate to yield accurate estimates.

	Method	uni	gauss	backg	impul	g-d-dec	l-d-dec	cut	oclsion	Mean
	P2P-Syn (Syn → Corrupted syn)									
	Source only	28.28	32.24	6.99	15.91	17.21	10.68	5.95	6.48	15.47 ± 30.14
EP	Point-TTA [12] (ICCV'23)	24.96	27.57	6.46	14.91	23.15	9.43	5.30	6.18	14.75 ± 29.55 (-0.72)
EP	Purge-Gate [38] (ICCV'25)	27.71	31.78	4.41	14.19	16.23	7.29	4.30	4.46	13.79 ± 28.35 (-1.68)
EP	LN [39] (WACV'26)	15.18	16.83	6.89	10.13	9.87	9.12	5.93	6.54	10.06 ± 22.43 (-5.41)
EP	PEA [20] (ICLR'26)	21.80	24.17	9.20	14.01	12.40	11.84	8.40	8.84	13.83 ± 27.66 (-1.64)
CT	LN [39] (WACV'26)	15.29	16.95	6.78	9.80	10.09	8.81	5.72	6.34	9.97 ± 22.38(-5.50)
CT	PEA [20] (ICLR'26)	22.93	25.46	7.93	13.67	15.46	10.98	7.39	8.05	13.98 ± 27.32(-1.49)
	P2ILReg-Syn (Syn → Corrupted syn)									
	Source only	88.31	104.24	23.49	82.09	3.35	2.23	2.85	3.07	38.70 ± 57.98
EP	Purge-Gate [38] (ICCV'25)	87.37	103.11	17.54	80.85	2.36	1.52	1.85	1.98	37.07 ± 57.57 (-1.63)
EP	LN [39] (WACV'26)	97.73	115.38	41.69	97.08	7.83	5.67	6.50	6.60	47.31 ± 62.61 (+8.61)
EP	PEA [20] (ICLR'26)	84.84	100.28	25.35	73.23	3.35	2.25	2.81	3.05	36.90 ± 56.48 (-1.80)
CT	LN [39] (WACV'26)	97.03	114.25	41.20	95.74	7.00	5.02	5.72	5.90	46.48 ± 61.80 (+7.78)
CT	PEA [20] (ICLR'26)	84.49	99.59	25.50	73.14	3.35	2.19	2.81	3.04	36.77 ± 56.20 (-1.93)

Table 1. Mean RMSE ± standard deviation (mm) of modified TTA methods on P2P-Syn and P2ILReg-Syn datasets. Continual (CT) and episodic (EP) denote the statistical estimation strategies.

		P2P-Syn Syn → Corrupted syn		**P2ILReg-Syn** Syn → Corrupted syn		**P2ILReg-Real** Syn → Real	
	Method	RRE (°)↓	RTE (mm)↓	RRE (°)↓	RTE (mm)↓	CD (mm)↓	Dice↑
	Source only	21.34 ± 46.87	14.81 ± 75.53	36.13 ± 56.33	24.25 ± 39.68	4.26 ± 1.82	71.87 ± 12.61
EP	Point-TTA [12] (ICCV'23)	19.58 ± 44.68	13.79 ± 69.18	–	–	–	–
EP	Purge-Gate [38] (ICCV'25)	18.89 ± 44.26	12.50 ± 65.60	34.69 ± 56.38	22.90 ± 38.85	4.17 ± 1.57	72.97 ± 10.85
EP	LN [39] (WACV'26)	12.83 ± 34.61	11.43 ± 64.21	43.09 ± 59.60	30.11 ± 44.75	5.34 ± 2.41	67.50 ± 14.51
EP	PEA [20] (ICLR'26)	18.08 ± 42.47	13.65 ± 68.48	34.19 ± 54.78	22.90 ± 39.00	4.27 ± 1.78	72.85 ± 12.18
CT	LN [39] (WACV'26)	12.80 ± 34.60	11.30 ± 64.07	42.61 ± 59.27	28.94 ± 43.57	5.59 ± 2.80	65.52 ± 15.17
CT	PEA [20] (ICLR'26)	18.12 ± 41.70	14.29 ± 69.48	34.20 ± 54.73	22.48 ± 38.29	4.20 ± 1.45	72.18 ± 11.42

Table 2. Mean RRE(°) and RTE(mm) ± standard deviation on synthetic datasets, and CD(mm) and Dice ± standard deviation on the real dataset.

Asymmetric adaptation: Normalization and input adaptation use statistics and embeddings that can be computed separately for each cloud. Model adaptation updates the shared encoder, so Point-TTA remains symmetric. Table 3 compares both regimes on P2ILReg, with symmetric adaptation computing shared statistics over both clouds and adapting both. For input adaptation, Purge-Gate and PEA are better in the asymmetric regime on both splits. For normalization adaptation, symmetric adaptation is better than asymmetric, but only improves the source model on P2ILReg-Real, not P2ILReg-Syn (38.79 vs 38.70 mm).
t-SNE visualizations: Fig. 2 suggests a limited adaptation, yielding a marginal shift toward the source mean. Purge-Gate only prunes points and cannot align them (Fig. 2). For PEA and LN, suboptimal alignment could come from inaccurate target statistic estimation. To address this, input adaptation methods could be extended with progressive alignment or purge across the encoder and decoder. Moreover, since PEA or LN accumulated statistics can become outdated, continual estimation methods [8,15] could improve them.
Computational cost: On the P2P dataset, the source-only average time per sample is 162 ms, and it is 184/189/414 ms for LN/PEA/Purge-Gate, compared to 1226 ms for Point-TTA. Peak GPU memory for source-only and the backpropagation-free methods stays at 1128–1187 MB, but reaches 5370 MB for Point-TTA. Since registration adaptation must be lightweight, this efficiency is decisive, and Point-TTA is less suited for real-time surgery.

		P2ILReg-Syn Syn → Corrupted syn		P2ILReg-Real Syn → Real			
		Symmetric	Asymmetric	Symmetric		Asymmetric	
	Method	RMSE (mm)↓	RMSE (mm)↓	CD (mm)↓	Dice↑	CD (mm)↓	Dice↑
	Source only	38.70 ± 57.98	38.70 ± 57.98	4.26 ± 1.82	71.87 ± 12.61	4.26 ± 1.82	71.87 ± 12.61
EP	Purge-Gate [38] (ICCV'25)	37.39 ± 58.19	37.07 ± 57.57	4.34 ± 1.87	71.51 ± 12.51	4.17 ± 1.57	72.97 ± 10.85
EP	LN [39] (WACV'26)	38.79 ± 58.27	47.31 ± 62.61	4.22 ± 1.67	72.21 ± 11.90	5.59 ± 2.80	65.52 ± 15.17
EP	PEA [20] (ICLR'26)	38.04 ± 57.42	36.90 ± 56.48	4.43 ± 1.94	72.21 ± 12.39	4.27 ± 1.78	72.85 ± 12.18

Table 3. RMSE (mm), CD (mm) and Dice ± standard deviation on P2ILReg datasets, under symmetric vs. asymmetric adaptation regimes.

5 Conclusion

Domain shift between synthetic training data and intraoperative reconstructions degrades laparoscopic registration. We analyzed and modified for registration four TTA methods across model, normalization, and input adaptation families. On corrupted synthetic and real data, input adaptation reduces registration errors, remaining the most robust family, with smaller and mixed gains on real data. Conversely, normalization adaptation only helps P2P, while model adaptation introduces computational overhead. This suggests that the TTA input adaptation family improves registration and could be further explored.

Acknowledgments. This work was supported in part by INOVAIT, the Government of Canada's Strategic Innovation Fund, and the Natural Sciences and Engineering Research Council of Canada. The Digital Research Alliance of Canada provided the computing resources. We thank Sébastien Delorme and his team at Scopia Surgical for contributing to the problem formulation.

Disclosure of Interests. The authors declare that they have no competing interests related to the content of this paper.

References

1. Alet, F., Bauza, M., Kawaguchi, K., et al.: Tailoring: encoding inductive biases by optimizing unsupervised objectives at prediction time. In: NeurIPS (2021)
2. Bahri, A., Yazdanpanah, M., Dastani, S., et al.: SMART-PC: Skeletal Model Adaptation for Robust Test-Time Training in Point Clouds. In: ICML (2025)
3. Bahri, A., Yazdanpanah, M., Noori, M., et al.: Test-Time Adaptation in Point Clouds: Leveraging Sampling Variation with Weight Averaging. In: WACV (2025)
4. Barrow, H.G., M, T.J., C, B.R., et al.: Parametric correspondence and chamfer matching : two new techniques for image matching. SRI internalional (1977)
5. Dastmalchi, H., An, A., Cheraghian, A., et al.: Test-Time Adaptation of 3D Point Clouds via Denoising Diffusion Models. In: WACV (2025)
6. Fang, Y., Yap, P.T., Lin, W., Zhu, H., Liu, M.: Source-free unsupervised domain adaptation: A survey. Neural Networks (2024)
7. Finn, C., Abbeel, P., Levine, S.: Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In: PMLR (2017)
8. Gao, Y., Xia, S., Nie, S., et al.: APCoTTA: Continual Test-Time Adaptation for Semantic Segmentation of Airborne LiDAR Point Clouds. In: ISPRS (2026)
9. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised Representation Learning by Predicting Image Rotations. In: ICLR (2018)

10. Gong, T., Jeong, J., Kim, T., et al.: Note: robust continual test-time adaptation against temporal correlation. In: NIPS (2022)
11. Guichemerre, A., Belharbi, S., Mayet, T., et al.: Source-Free Domain Adaptation of Weakly-Supervised Object Localization Models for Histology. In: CVPRw (2024)
12. Hatem, A., Qian, Y., Wang, Y.: Point-TTA: Test-Time Adaptation for Point Cloud Registration Using Multitask Meta-Auxiliary Learning. In: ICCV (2023)
13. Hu, X., Uzunbas, G., Chen, S., et al.: MixNorm: Test-Time Adaptation Through Online Normalization Estimation (2021)
14. Huang, B., Yang, X., Jia, F.: A Landmark-Free 3D–2D Rigid Liver Registration via Point Cloud Matching for Laparoscopic Surgery. HTL (2025)
15. Jiang, J., Zhou, Q., Li, Y., et al.: PCoTTA: Continual Test-Time Adaptation for Multi-Task Point Cloud Understanding. In: NeurIPS (2024)
16. Jiang, L., Ma, R., Gu, L., Wang, Z., Zuo, X., Wang, Y.: PointMAC: Meta-Learned Adaptation for Robust Test-Time Point Cloud Completion. In: NeurIPS (2025)
17. Kojima, T., Matsuo, Y., Iwasawa, Y.: Robustifying Vision Transformer without Retraining from Scratch by Test-Time Class-Conditional Feature Alignment (2022)
18. Liang, J., He, R., Tan, T.: A Comprehensive Survey on Test-Time Adaptation Under Distribution Shifts. IJCV (2025)
19. Ma, L., Huang, T., Wang, J., et al.: Visualization, registration and tracking techniques for augmented reality guided surgery: a review. PMB (2023)
20. Ma, X., Kwon, Y.D., Zhou, P., Ma, D.: Architecture-agnostic test-time adaptation via backprop-free embedding alignment. In: ICLR (2026)
21. Pan, S.J., Yang, Q.: A Survey on Transfer Learning. IEEE TKDE (2010)
22. Pang, Y., Wang, W., Tay, F.E.H., et al.: Masked Autoencoders for Point Cloud Self-supervised Learning. In: ECCV (2022)
23. Saltori, C., Krivosheev, E., Lathuiliére, S., et al.: Gipso: Geometrically informed propagation for online adaptation in 3d lidar segmentation. In: ECCV (2022)
24. Sharafi, M., Belharbi, S., Zeeshan, M.O., et al.: Personalized Feature Translation for Expression Recognition: An Efficient Source-Free Domain Adaptation Method. In: ICLR (2026)
25. Shim, H., Kim, C., Yang, E.: CloudFixer: Test-Time Adaptation for 3D Point Clouds via Diffusion-Guided Geometric Transformation. In: ECCV (2024)
26. Sun, H., Ke, Q., Cheng, M., et al.: Point-Cache: Test-time Dynamic and Hierarchical Cache for Robust and Generalizable Point Cloud Analysis. In: CVPR (2025)
27. Sun, J., Zhang, Q., Kailkhura, B., Yu, Z., Xiao, C., Mao, Z.M.: Benchmarking Robustness of 3D Point Cloud Recognition Against Common Corruptions (2022)
28. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully Test-time Adaptation by Entropy Minimization. In: ICLR (2021)
29. Wang, X., Tsvetkov, Y., Ruder, S., et al.: Efficient Test Time Adapter Ensembling for Low-resource Language Varieties. In: EMNLP (2021)
30. Wang, Y., Cheraghian, A., Hayder, Z., et al.: Backpropagation-free Network for 3D Test-time Adaptation. In: CVPR (2024)
31. Wei, X., Yang, Q., Fang, Y., et al.: 3D Test-time Adaptation via Graph Spectral Driven Point Shift. In: ICCV (2025)
32. Welford, B.P.: Note on a method for calculating corrected sums of squares and products. Technometrics (1962)
33. Xiao, Z., Snoek, C.G.M.: Beyond Model Adaptation at Test Time: A Survey (2024)
34. Yang, J., Zhang, C., Wang, Z., et al.: 3D Registration in 30 Years: A Survey (2024)
35. Yang, Z., Heiselman, J.S., Han, C., et al.: Resolving the Ambiguity of Complete-to-Partial Point Cloud Registration for Image-Guided Liver Surgery with Patches-to-Partial Matching. IEEE JBHI (2026)

36. Yang, Z., Simon, R., Linte, C.A.: Learning Feature Descriptors for Pre- and Intra-operative Point Cloud Matching for Laparoscopic Liver Registration. IJCARS (2023)
37. Yao, R., Du, S., Cui, W., Tang, C., Yang, C.: PARE-Net: Position-Aware Rotation-Equivariant Networks for Robust Point Cloud Registration. In: ECCV (2024)
38. Yazdanpanah, M., Bahri, A., Noori, M., et al.: Purge-Gate: Backpropagation-Free Test-Time Adaptation for Point Clouds Classification via Token purging. In: ICCV (2025)
39. Yazdanpanah, M., Bahri, A., Noori, M., et al.: Revisiting layer normalization for point cloud test time adaptation. In: WACV (2026)
40. Zhang, Y.F., Wang, J., Liang, J., et al.: Domain-specific risk minimization for domain generalization. In: KDD (2023)
41. Zhang, Y.X., Gui, J., Yu, B., Cong, X., Gong, X., Tao, W., Tao, D.: Deep Learning-Based Point Cloud Registration: A Comprehensive Survey and Taxonomy. IJCV (2026)
42. Zhou, J., Gao, B., Wang, K., et al.: Landmark-Free Preoperative-to-Intraoperative Registration in Laparoscopic Liver Resection. IEEE TMI (2025)