CoRe-DA: Contrastive Regression for Unsupervised Domain Adaptation in Surgical Skill Assessment

Dimitrios Anastasiou[1,2](✉), Razvan Caramalau[1], Jialang Xu[1,2], Runlong He[1,2], Freweini Tesfai[4], Matthew Boal[4], Nader Francis[4], Danail Stoyanov[1,3], and Evangelos B. Mazomenos[1,2](✉)

[1] UCL Hawkes Institute, University College London, UK
[2] Dept of Medical Physics & Biomedical Engineering, University College London, UK
[3] Dept of Computer Science, University College London, UK
[4] The Griffin Institute, Northwick Park Institute for Medical Research, London, UK
{dimitrios.anastasiou.21;e.mazomenos}@ucl.ac.uk

Abstract. Vision-based surgical skill assessment (SSA) enables objective and scalable evaluation of operative performance. Progress in this field is constrained by the high cost and time demands for manual annotation of quantitative skill scores, as well as the poor generalization of existing regression models to new surgical tasks and environments. Meanwhile, appreciable volumes of unlabeled video data are now available, motivating the development of unsupervised domain adaptation (UDA) methods for SSA. We introduce the first benchmark for UDA in SSA regression, spanning four datasets across dry-lab and clinical settings as well as open and robotic surgery. We evaluate eight representative models under challenging domain shifts and propose CoRe-DA, a novel contrastive regression–based adaptation framework. Our method learns domain-invariant representations through relative-score supervision and target-domain self-training. Comprehensive experiments across two UDA settings show that CoRe-DA is superior to state-of-the-art methods, achieving Spearman Correlation Coefficients of 0.46 and 0.41 on dry-lab and clinical target datasets, respectively, without using any labeled target data for training. Overall, CoRe-DA enables scalable SSA with reliable cross-domain generalization, where existing methods underperform. Our code and datasets are available at CoRe-DA.

Keywords: Surgical Skill Assessment · Unsupervised Domain Adaptation · Contrastive Regression · Surgical Video Analysis

1 Introduction

Automated surgical skill assessment (SSA) has emerged as a promising approach for objective and scalable evaluation of surgical performance [31], yet its adoption is constrained by the scarcity of labeled data [2,28]. SSA is typically formulated as a regression task, where annotations are usually produced using standardized

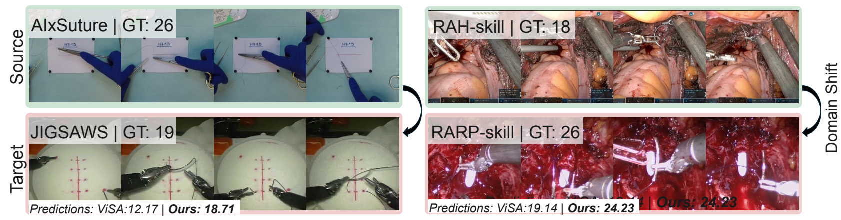

Fig. 1. Sample sequences illustrating substantial domain shifts for surgical skill assessment (SSA), with target predictions from ViSA (SOTA) and CoRe-DA (ours).

frameworks such as the Objective Structured Assessment of Technical Skill (OSATS) [15], which rates multiple skill components on 5-point Likert scales. Producing such labels is labor-intensive, as raters (expert surgeons) must repeatedly review videos and coordinate to reach consensus, leading to a substantial annotation cost. Meanwhile, although current state-of-the-art (SOTA) SSA regression models report strong performance on dry-lab and clinical datasets [3, 9, 13, 14], they remain largely domain-specific and generalize poorly to new surgical procedures and environments (see Table 1). This makes large-scale deployment challenging under realistic annotation budgets. In contrast, the volume of unlabeled surgical recordings has been progressively increasing [24]. This motivates us to address the following question: *Can we learn an SSA regression model that generalizes reliably to a* ***target*** *domain when trained only with labeled* ***source*** *domain data and* ***unlabeled*** *target data? In other words, how far can we go with the data and labels already available?*

Unsupervised Domain Adaptation (UDA) finds direct application in this problem. UDA leverages labeled data from a source domain and unlabeled data from a target domain to improve model generalization in the target domain [26]. Homologous to our regression task, prior work on UDA includes RSD [7], which aligns domains using orthogonal bases of their representation spaces, and DARE-GRAM [16], which performs adaptation by matching inverse Gram matrices of feature representations. Prior work on UDA in SSA is limited [10,23] and formulates SSA as binary classification, which fundamentally oversimplifies the task and limits the granularity of feedback that can be provided to surgeons. In contrast, we formulate SSA prediction as regression (score $\in \mathbb{R}$), enabling fine-grained and more informative assessment of surgical performance. Moreover, [23] relies on robotic kinematic data, which are typically unavailable, and is evaluated only on simulated environments, while [10] is exclusive to ophthalmology.

We extend the SOTA by introducing the first formulation of UDA for SSA regression, and establishing a benchmark comprising two UDA settings across four datasets spanning open, robotic dry-lab tasks and clinical robotic surgery (see Fig. 1), evaluated with eight representative models. These benchmarks present substantial domain shifts due to differences in procedures, acquisition setups, surgical instruments, and visual appearance, making adaptation particularly challenging. Additionally, we propose **CoRe-DA**, a novel contrastive regression–based UDA framework for SSA which, to the best of our knowledge, is the first to explore contrastive regression as an adaptation strategy. Our formulation is motivated by the observation that standardized SSA tools (*e.g.,* OSATS) assess domain-agnostic performance criteria across procedures and environments, yet regression models struggle to directly estimate performance scores in cross-domain settings. Instead, differences in surgical actions defining skill across videos provide a more reliable signal for learning domain-invariant representations and guiding adaptation. Unlike adversarial [30] or feature-alignment approaches [7, 16], CoRe-DA learns domain-invariant representations by contrasting pairs of videos drawn from both source and target domains. It is trained using absolute and relative score regression losses on source data and regularized through a consistency constraint between the regression heads. In addition, CoRe-DA incorporates self-training on target samples, enabled by pseudo-labels, which is pivotal for driving adaptation and stabilizing regression under severe domain shift. Our main contributions are:

1. We formulate UDA for SSA regression and introduce a comprehensive benchmark comprising two challenging UDA settings across four diverse datasets and eight representative baseline models.
2. We propose CoRe-DA, a novel contrastive regression–based UDA framework for SSA. CoRe-DA represents the first application of contrastive regression as an adaptation strategy in UDA.
3. Extensive experiments show that CoRe-DA clearly outperforms SOTA methods, achieving improvements in Spearman's Correlation Coefficient (SCC) of +0.13 and +0.26, and reductions in Mean Absolute Error of 0.32 and 0.49 on dry-lab and clinical robotic surgery target datasets, respectively.

2 Methods

2.1 Preliminaries

UDA uses labeled data from a source domain and unlabeled data from a target domain to improve model generalization in the target domain. We are given a labeled source dataset $\mathcal{D}_S = \{(x_S^i, y_S^i)\}_{i=1}^{N_S}$ and an unlabeled target dataset $\mathcal{D}_T = \{x_T^i\}_{i=1}^{N_T}$, with a common label space. Our goal is to learn a regression model that generalizes to the target domain under distribution shift.

2.2 Framework (CoRe-DA)

Overview: CoRe-DA, shown in Fig. 2, samples batches of triplets (x_S, x_E, x_T) during training, consisting of two (different) labeled source videos, termed *source* and *exemplar*, and an unlabeled *target* video. A shared encoder $\mathcal{F}$ maps these inputs to feature representations, processed by a relative regressor $\mathcal{R}_{\text{rel}}$ to produce score differences between video pairs (relative score), and an absolute regressor $\mathcal{R}_{\text{abs}}$, to produce individual (absolute) scores. CoRe-DA learns domain-invariant representations via contrastive regression by comparing source–exemplar and

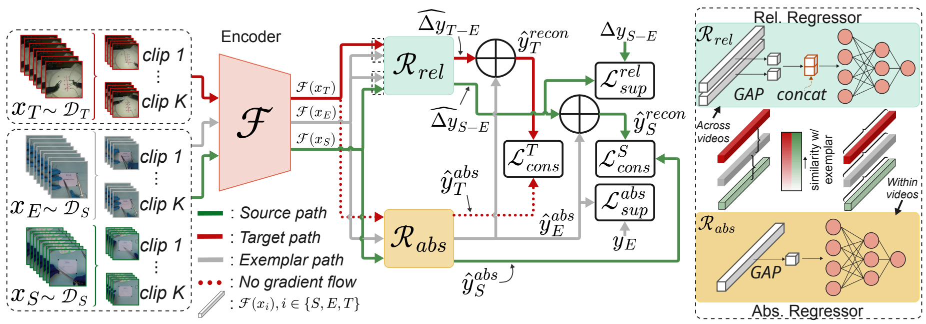

Fig. 2. Overview of the proposed CoRe-DA framework. Training uses source–exemplar–target triplets processed by a shared encoder and absolute and relative regression heads. Domain adaptation is driven by contrastive regression between source–exemplar and target–exemplar pairs and pseudo-label self-training.

target–exemplar pairs. Training uses supervised losses on source predictions together with a self-training strategy on target samples to further drive adaptation.
Clip Sampling and Model Architecture: To promote temporal feature diversity, we employ stochastic clip sampling during training as in [22, 27]. Given a video of L frames, we divide it into K non-overlapping equal-length segments and randomly sample a clip of l consecutive frames from each, producing a tensor $x \in \mathbb{R}^{(K \cdot l) \times c \times h \times w}$, with c, h, and w denoting channels, height, and width. We use x to refer to both the full video and its sampled representation. To align with prior UDA work [8], we adopt I3D [6] as the feature encoder $\mathcal{F} : \mathbb{R}^{(K \cdot l) \times c \times h \times w} \to \mathbb{R}^{K \times d}$, where d is the feature dimension. The absolute regressor $\mathcal{R}_{\text{abs}} : \mathbb{R}^{K \times d} \to \mathbb{R}$ is implemented via global average pooling (GAP) over the temporal (K-clip) dimension, followed by a 3-layer MLP. Following [4], the relative regressor $\mathcal{R}_{\text{rel}}$ operates on concatenated pooled features of video pairs x_i and x_j (*i.e.*, concat[GAP($\mathcal{F}(x_i)$); GAP($\mathcal{F}(x_j)$)]), producing a mapping $\mathcal{R}_{\text{rel}} : \mathbb{R}^{K \times 2d} \to \mathbb{R}$. Similarly to $\mathcal{R}_{\text{abs}}$, $\mathcal{R}_{\text{rel}}$ employs a 3-layer MLP.

Contrastive Regression-based Domain Adaption: For a triplet (x_S, x_E, x_T) we compute absolute and relative score predictions as:

$$\hat{y}_i^{\text{abs}} = \mathcal{R}_{\text{abs}}(\mathcal{F}(x_i)) , \quad i \in \{S, E, T\}, \tag{1}$$

$$\widehat{\Delta y}_{j-E} = \mathcal{R}_{\text{rel}}(\mathcal{F}(x_j), \mathcal{F}(x_E)) , \quad j \in \{S, T\}. \tag{2}$$

Considering $\widehat{\Delta y}_{i-j} \triangleq \hat{y}_i - \hat{y}_j$, we *reconstruct* absolute predictions from relative ones as $\hat{y}_j^{\text{recon}} = \widehat{\Delta y}_{j-E} + \hat{y}_E^{\text{abs}}, j \in \{S, T\}$.

CoRe-DA is trained on labeled source samples with two supervised losses: $\mathcal{L}_{\text{sup}}^{\text{rel}} = \frac{1}{B} \sum_{i=1}^{B} (\widehat{\Delta y}_{S-E}^{(i)} - \Delta y_{S-E}^{(i)})^2$ and $\mathcal{L}_{\text{sup}}^{\text{abs}} = \frac{1}{B} \sum_{i=1}^{B} (\hat{y}_E^{\text{abs},(i)} - y_E^{(i)})^2$, where $\Delta y_{S-E} = y_S - y_E$, y_S and y_E are the *source* and *exemplar* score labels, and B is the batch size. $\mathcal{L}_{\text{sup}}^{\text{abs}}$ encourages $\mathcal{F}$ to produce score-discriminative representations [19], and ensures that the score scale is preserved. Importantly, $\mathcal{L}_{\text{sup}}^{\text{rel}}$ pro-

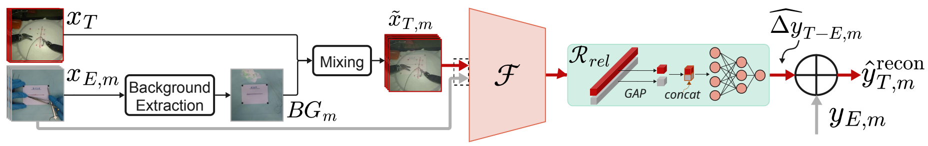

Fig. 3. Target-domain testing: Target videos are mixed with exemplar backgrounds, and predictions are reconstructed from relative scores computed across multiple pairs.

motes domain-invariant feature learning through contrastive supervision on relative predictions. To avoid prediction drift between the two regressors, we further enforce the following consistency constraint $\mathcal{L}^{S}_{\text{cons}} = \frac{1}{B}\sum_{i=1}^{B}(\hat{y}_S^{recon,(i)} - \hat{y}_S^{abs,(i)})^2$.

Target-domain adaptation is driven by a self-training objective aligning absolute predictions from $\mathcal{R}_{\text{abs}}$ with reconstructed predictions via $\mathcal{R}_{\text{rel}}$. Rather than relying on multiple augmented views of the same input, as in standard consistency-based methods [21], our method derives supervision from two complementary prediction pathways (absolute and relative). Inspired by [20], we stop the gradient flow through the absolute prediction branch when computing this loss to prevent model collapse. This process can also be interpreted as $\mathcal{R}_{\text{abs}}$ generating pseudo-labels for target-domain supervision. The resulting target-domain consistency loss is defined as $\mathcal{L}^{T}_{\text{cons}} = \frac{1}{B}\sum_{i=1}^{B}(\hat{y}_T^{recon,(i)} - \text{stopgrad}(\hat{y}_T^{abs,(i)}))^2$. The total loss is given by $\mathcal{L}_{\text{total}} = \alpha(\mathcal{L}^{\text{rel}}_{\text{sup}} + \mathcal{L}^{\text{abs}}_{\text{sup}}) + \beta\mathcal{L}^{S}_{\text{cons}} + \gamma\mathcal{L}^{T}_{\text{cons}}$, where α, β, and γ are scaling hyperparameters.

Target Domain Testing: As shown in Fig. 3, we employ $\mathcal{R}_{rel}$ using multiple exemplars as in [29]. To improve robustness to domain shift, we further introduce a single-pass test-time transformation, different from stochastic test-time augmentation, in which exemplar backgrounds are mixed with target videos [18]. Given M exemplar videos $\{x_{E,m}\}_{m=1}^{M}$, sampled uniformly from $\mathcal{D}_S$ and stratified by score label, we generate a single background frame for each using temporal median filtering as in [18], resulting in $\{BG_m\}_{m=1}^{M}$. Each background is then mixed with all frames of the target video x_T, as $\tilde{x}_{T,m} = (1-\lambda)\,x_T + \lambda\,BG_m$, where $\lambda \in [0,1]$ controls the mixing ratio. For each pair $(\tilde{x}_{T,m}, x_{E,m})$, we produce relative score predictions using Eq. 2 and reconstruct target scores as $\hat{y}^{\text{recon}}_{T,m} = \widehat{\Delta y}_{T-E,m} + y_{E,m}$, where $y_{E,m}$ is the exemplar label. The final target prediction is the average over all exemplars: $\bar{\hat{y}}_T = \frac{1}{M}\sum_{m=1}^{M}\hat{y}^{\text{recon}}_{T,m}$.

3 Experiments

Benchmark Datasets and UDA Settings: We benchmark UDA for SSA on two settings across four datasets: AIxSuture [12], JIGSAWS [1], RARP-skill [11], and RAH-skill [11]. AIxSuture is a dry-lab open-surgery dataset of 314 suturing videos annotated with OSATS across 8 components (each scored 1-5). JIGSAWS comprises 103 dry-lab robotic videos of suturing, needle passing, and knot tying, with 6-component OSATS labels. RARP-skill is a 33-video subset

Table 1. UDA results on our two settings (labeled source + unlabeled target), averaged over four runs with different random seeds. Top two results are in **bold** and <u>underlined</u>.

Method	Target Adaptat.	AIxSuture→JIGSAWS		RAH-skill→RARP-skill	
		SCC (↑)	MAE (↓)	SCC (↑)	MAE (↓)
Source-Only	✗	0.20 (±0.15)	4.84 (±0.36)	0.03 (±0.04)	2.89 (±0.32)
Contra-Sformer [3]	✗	0.14 (±0.11)	7.45 (±1.14)	0.03 (±0.06)	2.66 (±0.39)
ViSA [13]	✗	0.07 (±0.26)	6.12 (±0.70)	0.06 (±0.06)	3.10 (±0.66)
MDD [30]	✓	0.12 (±0.05)	5.02 (±0.16)	<u>0.15</u> (±0.01)	3.02 (±0.48)
DARE-GRAM [16]	✓	0.29 (±0.04)	4.80 (±0.09)	0.14 (±0.05)	2.54 (±0.05)
RSD [7]	✓	0.30 (±0.02)	<u>4.61</u> (±0.10)	0.07 (±0.05)	<u>2.42</u> (±0.36)
CO^2A [8]	✓	<u>0.33</u> (±0.08)	5.04 (±0.72)	0.13 (±0.06)	3.60 (±0.28)
CoRe-DA (ours)	✓	**0.46** (±0.03)	**4.29** (±0.08)	**0.41** (±0.05)	**1.93** (±0.07)

of SAR-RARP50 [17] of dorsal venous complex suturing, annotated with the Modified Global Evaluative Assessment of Robotic Skills (M-GEARS: 6 components, each scored 1-5) [5]. RAH-skill contains 37 robotic hysterectomy videos of vaginal cuff closure, each with three suturing tasks, yielding 110 annotated clips with M-GEARS. Based on these datasets, we define two UDA settings under the constraints that $\mathcal{D}_S$ and $\mathcal{D}_T$ share an identical label space and that $\mathcal{D}_S$ is sufficiently large for supervised training [25]. We assume marginal distribution shift with invariant label-conditional distribution. The first uses $\mathcal{D}_S$ = AIxSuture and $\mathcal{D}_T$ = JIGSAWS, keeping their 6 overlapping OSATS components. The second uses $\mathcal{D}_S$ = RAH-skill and $\mathcal{D}_T$ = RARP-skill, with the same 6-components M-GEARS. The sum of all components is the final score label ($\in [6, 30]$).

Evaluation Protocol and Implementation Details: Following standard UDA protocols [7, 30], for each UDA setting, we use all **labeled** source data and all **unlabeled** target data for training, and test on all target data. In line with [3, 13], we report test performance with SCC and Mean Absolute Error (MAE). We initialize I3D with Kinetics weights, and set $d = 256$. Both regressors use 3-layer MLPs with output dimensions 256/128/1. Videos are downsampled to 1 fps and resized to 224×224. For training, we sample $l = 12$ frames from $K = 12$ segments. CoRe-DA is trained with Adam for 150 epochs using learning rates 10^{-5} for I3D and 5×10^{-5} for $\mathcal{R}_{\text{abs}}$ and $\mathcal{R}_{\text{rel}}$, batch size 16, and α=β=γ=1. For testing, we sample consecutive non-overlapping clips ($l = 12$) to cover the full video. We use $M = 10$ exemplars, and set $\lambda = 0.25$. CoRe-DA is implemented in PyTorch and trained on an RTX A6000 (48 GB).

4 Results

Comparison with the SOTA: We benchmark UDA performance for SSA and compare Core-DA against established UDA baselines: MDD [30], DARE-GRAM [16], RSD [7], and CO^2A [8]. We also include domain-specific SOTA SSA models, ViSA [13] and Contra-Sformer [3], and a Source-Only baseline

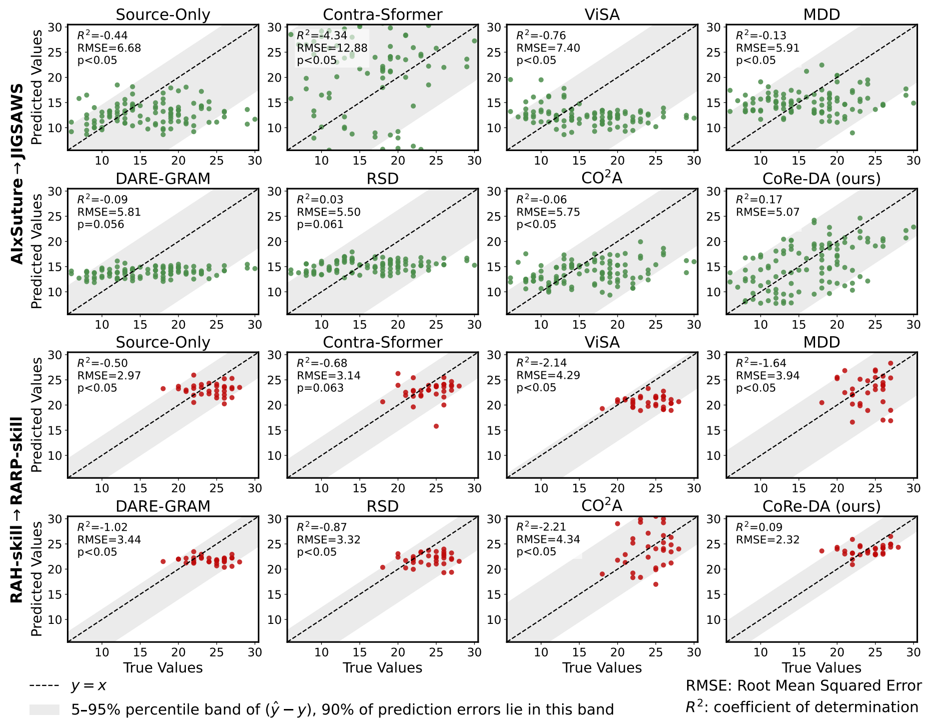

Fig. 4. Scatter plots of predicted vs true scores (run #1 displayed). Paired t-tests across all 4 runs show significant (or marginally) improvements for CoRe-DA against nearly all methods, with p-value ranges: Source-Only (0.02-0.06/$<$ 0.05), Contra-S. ($<$ 0.05/0.00-0.07), ViSA ($<$ 0.05/$<$ 0.05), MDD ($<$ 0.05/$<$ 0.05), DARE-G. (0.02-0.06/$<$ 0.05), RSD (0.04-0.16/$<$ 0.05), and CO^2A (0.04-0.08/$<$ 0.05) for JIGSAWS/RARP-skill, respectively. These trends are further supported by higher R^2 and lower RMSE.

(I3D+regressor as in [8]), all trained on labeled source data and evaluated on the target domain without adaptation. All methods are implemented from released code and adapted to our task; [10] is omitted, due to code unavailability. For fairness, we use I3D in all UDA methods, while keeping the original architectures for ViSA and Contra-Sformer. As shown in Table 1, most UDA methods outperform the Source-Only baseline, confirming the benefit of leveraging unlabeled target data. CoRe-DA yields the best performance, improving SCC by +0.13 and MAE by -0.32 in AIxSuture →JIGSAWS, and boosting SCC by +0.26 with a MAE reduction of -0.49 in RAH-skill→RARP-skill, compared to second-best baselines. Contra-Sformer and ViSA generalize poorly to target domains, reflecting the limitations of domain-specific SSA models under domain shift. CoRe-DA is the only method achieving consistently, with low standard deviation, strong performance across both metrics and settings (MAE of 4.29 is reasonable for the $[6, 30]$ JIGSAWS range). Fig. 4 shows scatter plots from a single run. AIxSuture→JIGSAWS proves particularly challenging due to the large

Table 2. 10-shot Semi-Supervised UDA results on AIxSuture→JIGSAWS, averaged over five labeled-target subsets.

Method	SCC (↑)	MAE (↓)
Sour.+Targ.	0.39 (±0.08)	4.50 (±0.42)
Contra-S. [3]	0.34 (±0.19)	5.20 (±0.84)
ViSA [13]	0.32 (±0.03)	4.48 (±0.27)
MDD [30]	0.36 (±0.05)	4.44 (±0.30)
DARE. [16]	0.40 (±0.07)	4.14 (±0.15)
RSD [7]	0.42 (±0.07)	4.40 (±0.17)
CO^2A [8]	0.46 (±0.05)	4.31 (±0.32)
CoRe-DA	**0.52** (±0.04)	**3.83** (±0.21)

Table 3. Ablation studies on AIxSuture→JIGSAWS. Results correspond to a single run using the same random seed.

$\mathcal{L}_{sup}^{rel}$	$\mathcal{L}_{sup}^{abs}$	$\mathcal{L}_{cons}^{S}$	$\mathcal{L}_{cons}^{T}$	stop grad	BG_m	SCC	MAE
✗	✓	✓	✓	✓	✓	0.38	4.89
✓	✗	✓	✓	✓	✓	0.17	5.64
✓	✓	✗	✓	✓	✓	0.44	4.57
✓	✓	✓	✗	✗	✓	0.29	4.75
✓	✓	✓	✓	✗	✓	0.33	4.66
✓	✓	✓	✓	✓	✗	0.43	4.57
✓	✓	✓	✓	✓	✓	**0.47**	**4.41**

domain gap and wider score range, with most methods collapsing predictions toward the mean. In contrast, CoRe-DA covers a substantially wider score range, preserves the score ordering, and has smaller error fluctuations, as reflected by a narrower percentile band. Similar trends are observed for RAH-skill→RARP-skill, where CoRe-DA produces more monotonic predictions with consistently lower errors. We further evaluate a 10-shot semi-supervised UDA setting on AIxSuture→JIGSAWS, where labels from 10 target videos are available during training and are incorporated with a supervised ℓ_2 loss. These samples are excluded from testing. In Table 2, CoRe-DA outperforms all other methods, demonstrating its practical value when only a small number of target domain annotations is available. Lastly, as SSA is post-operative rather than real-time, inference cost is less critical; CoRe-DA requires 1760.424 GFLOPs for $M = 10$, compared with 670.638 for Source-Only/MDD/DARE-GRAM/RSD, 2177.518 for Contra-Sformer, 912.435 for ViSA, and 889.276 for CO^2A.

Ablation Studies: We evaluate the impact of key CoRe-DA components in Table 3. Without $\mathcal{L}_{\text{sup}}^{\text{rel}}$ performance drops substantially, showing that relative score supervision is essential for learning domain-invariant representations. Removing $\mathcal{L}_{\text{sup}}^{\text{abs}}$ further degrades performance, as $\mathcal{R}_{\text{abs}}$ no longer learns the score scale reliably, affecting the quality of target pseudo-labels. $\mathcal{L}_{\text{cons}}^{S}$ has a small but consistent effect as the two regressors may gradually converge to similar solutions. Without $\mathcal{L}_{\text{cons}}^{T}$ target adaptation is effectively disabled, causing a substantial performance drop which highlights the importance of our self-training design for successful domain adaptation. Stopping gradient flow through the absolute branch is essential, as performance otherwise degrades likely due to model collapse. Test-time background mixing improves adaptation by partially aligning low-level visual statistics across domains; still, CoRe-DA outperforms all methods in Table 1 even without it. Lastly, increasing the number of exemplars from 2 to 12 monotonically improves SCC (0.43–0.47) and MAE (4.79–4.40), plateauing at 12; very small λ values have little effect, whereas $\lambda > 0.25$ distort target frames and degrade performance.

5 Conclusions

We present the first benchmark for UDA in SSA regression, consisting of two UDA settings and four datasets across dry-lab and clinical environments. We introduce CoRe-DA, a novel contrastive regression–based framework that learns domain-invariant representations via contrastive supervision and target self-training. CoRe-DA outperforms SOTA methods, achieving SCC of 0.46 and 0.41 and MAE of 4.29 and 1.93 on JIGSAWS and RARP-skill, respectively, without using any target labels. Our results demonstrate the potential of our method to enable scalable, video-based SSA in both dry-lab and clinical settings. Future work will explore incorporating uncertainty estimation to improve adaptation by filtering unreliable target pseudo-labels, disentangling individual domain shifts, and extending CoRe-DA to individual skill-component scoring.

Acknowledgments. This study was funded in whole, or in part, by the EPSRC [EP/R513143/1, EP/T517793/1, EP/S021930/1, EP/Z534754/1, UKRI145], the DSIT and the RAEng under the Chair in Emerging Technologies.

Disclosure of Interests. DS reports employment with Medtronic Ltd; board membership, consulting or advisory, and equity or stocks with Odin Medical Ltd, Panda Surgical Ltd and EnAcuity Ltd. The remaining authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahmidi, N., Tao, L., Sefati, S., Gao, Y., Lea, C., Haro, B.B., Zappella, L., Khudanpur, S., Vidal, R., Hager, G.D.: A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. IEEE Trans. Biomed. Eng. **64**(9), 2025–2041 (2017)
2. Anastasiou, D., Caramalau, R., Sirajudeen, N., Boal, M., Edwards, P., Collins, J., Kelly, J., Sridhar, A., Tran, M., Mumtaz, F., Pavithran, N., Francis, N., Stoyanov, D., Mazomenos, E.B.: Exploring pre-training across domains for few-shot surgical skill assessment. In: DEMI. pp. 212–222. Springer Nature Switzerland (2026)
3. Anastasiou, D., Jin, Y., Stoyanov, D., Mazomenos, E.: Keep your eye on the best: Contrastive regression transformer for skill assessment in robotic surgery. IEEE Robot. Autom. Lett. **8**(3), 1755–1762 (2023)
4. Bai, Y., Zhou, D., Zhang, S., Wang, J., Ding, E., Guan, Y., Long, Y., Wang, J.: Action quality assessment with temporal parsing transformer. In: ECCV. pp. 422–438. Lecture Notes in Computer Science, Springer (2022)
5. Boal, M.W.E., Afzal, A., Gorard, J., Shah, A., Tesfai, F., Ghamrawi, W., Tutton, M., Ahmad, J., Selvasekar, C., Khan, J., Francis, N.K.: Development and evaluation of a societal core robotic surgery accreditation curriculum for the uk. J. Robot. Surg. **18**(1), 305 (2024)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the Kinetics dataset. In: CVPR. pp. 4724–4733 (2017)
7. Chen, X., Wang, S., Wang, J., Long, M.: Representation subspace distance for domain adaptation regression. In: ICML. vol. 139, pp. 1749–1759. PMLR (2021)

8. Turrisi da Costa, V.G., Zara, G., Rota, P., Oliveira-Santos, T., Sebe, N., Murino, V., Ricci, E.: Dual-head contrastive domain adaptation for video action recognition. In: WACV. pp. 2234–2243 (2022)
9. Ding, X., Xu, X., Li, X.: Sedskill: Surgical events driven method for skill assessment from thoracoscopic surgical videos. In: MICCAI. pp. 35–45. LNCS, Springer Nature Switzerland (2023)
10. Gong, Z., Wan, B., Paranjape, J.N., Sikder, S., Patel, V.M., Vedula, S.S.: Evaluating the generalizability of video-based assessment of intraoperative surgical skill in capsulorhexis. Int. J. Comput. Assist. Radiol. Surg. **20**(11), 2231–2239 (2025)
11. He, R., Tesfai, F.M., Boal, M.W.E., Sirajudeen, N., Anastasiou, D., Xu, J., Hoque, M.I., Edwards, P.J., Kelly, J.D., Sridhar, A., Kadkhodamohammadi, A., Chandrasekaran, D., Clarkson, M.J., Stoyanov, D., Francis, N., Mazomenos, E.B.: Surgfusion-net: Diversified adaptive multimodal fusion network for surgical skill assessment. arXiv preprint arXiv:2603.00108 (2026)
12. Hoffmann, H., Funke, I., Peters, P., Venkatesh, D.K., Egger, J., Rivoir, D., Röhrig, R., Hölzle, F., Bodenstedt, S., Willemer, M., Speidel, S., Puladi, B.: Aixsuture: vision-based assessment of open suturing skills. Int. J. Comput. Assist. Radiol. Surg. **19**, 1045–1052 (2024)
13. Li, Z., Gu, L., Wang, W., Nakamura, R., Sato, Y.: Surgical skill assessment via video semantic aggregation. In: MICCAI. pp. 488–498. LNCS, Springer Nature Switzerland (2022)
14. Liu, D., Li, Q., Jiang, T., Wang, Y., Miao, R., Shan, F., Li, Z.: Towards unified surgical skill assessment. In: CVPR. pp. 9522–9531 (2021)
15. Martin, J.A., Regehr, G., Reznick, R., Macrae, H., Murnaghan, J., Hutchison, C., Brown, M.: Objective structured assessment of technical skill (osats) for surgical residents. Br. J. Surg. **84**, 273–278 (1997)
16. Nejjar, I., Wang, Q., Fink, O.: Dare-gram : Unsupervised domain adaptation regression by aligning inversed gram matrices. In: CVPR. pp. 11744–11754 (2023)
17. Psychogyios, D., Colleoni, E., Amsterdam, B.V., Li, C.Y., Huang, S.Y., Li, Y., Jia, F.: Sar-rarp50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. arXiv:2401.00496 (2023)
18. Sahoo, A., Shah, R., Panda, R., Saenko, K., Das, A.: Contrast and mix: Temporal contrastive video domain adaptation with background mixing. In: NeurIPS. pp. 23386–23400 (2021)
19. Samarasinghe, S., Rizve, M.N., Kardan, N., Shah, M.: Cdfsl-v: Cross-domain few-shot learning for videos. In: ICCV. pp. 11643–11652 (2023)
20. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS (2017)
21. Van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. Machine learning **109**(2), 373–440 (2020)
22. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks for action recognition in videos. IEEE Trans. Pattern Anal. Mach. Intell. **41**(11), 2740–2755 (2019)
23. Wang, Z., Mariani, A., Menciassi, A., De Momi, E., Majewicz Fey, A.: Uncertainty-aware self-supervised learning for cross-domain technical skill assessment in robot-assisted surgery. IEEE Trans. Med. Robot. Bionics **5**(2), 301–312 (2023)
24. Wei, J., Xiao, Z., Sun, D., Gong, L., Yang, Z., Liu, Z., Wu, J.: Surgbench: A unified large-scale benchmark for surgical video analysis. arXiv preprint arXiv:2506.07603 (2025)

25. Wilson, G., Cook, D.J.: A survey of unsupervised deep domain adaptation. ACM Trans. Intell. Syst. Technol. **11**(5), 1–46 (2020)
26. Xu, Y., Cao, H., Xie, L., Li, X.l., Chen, Z., Yang, J.: Video unsupervised domain adaptation with deep learning: A comprehensive survey. ACM Comput. Surv. **56**(12), 1–36 (2024)
27. Xu, Y., Yang, J., Zhou, Y., Chen, Z., Wu, M., Li, X.: Augmenting and aligning snippets for few-shot video domain adaptation. In: ICCV. pp. 13445–13456 (2023)
28. Yanik, E., Schwaitzberg, S., Yang, G., Intes, X., Norfleet, J., Hackett, M., De, S.: One-shot skill assessment in high-stakes domains with limited data via meta learning. Comput. in Biol. and Med. **174**, 108470 (2024)
29. Yu, X., Rao, Y., Zhao, W., Lu, J., Zhou, J.: Group-aware contrastive regression for action quality assessment. In: ICCV. pp. 7919–7928 (2021)
30. Zhang, Y., Liu, T., Long, M., Jordan, M.: Bridging theory and algorithm for domain adaptation. In: ICML. vol. 97, pp. 7404–7413. PMLR (2019)
31. Zia, A., Essa, I.: Automated surgical skill assessment in rmis training. Int. J. Comput. Assist. Radiol. Surg. **13**(5), 731–739 (2018)