

Peer Review

Review of: "Deep Learning in Medical Image Registration: Introduction and Survey"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

This paper is a tutorial and a survey of medical image registration (MIR). It sets out to introduce the topic from the beginning and then review recent work. It opens in Section 1 with the origin of the word "registration" in dictionaries and in the printing industry. In Section 2.2 (Equation 1), it gives its own definition of registration as finding a space in which corresponding elements are close. Section 3 then walks through a numeric example of a displacement field and resampling on a 2x2-pixel image.

The survey part is described in Section 4 ("Review criterion"). The authors searched Scopus with the keywords "medical image registration" for open-access papers in English published between 2021 and 2022. They retrieved 270 records, excluded 38 on the abstract as irrelevant, and excluded 41 because no open-access version was found as of December 2022. Of the remaining 191 papers, 96 were reviewed, plus 10 papers published before 2021. Table 3 states the research question: what was the research pipeline of MIR in 2021 and 2022. Its sub-questions cover proposed approaches, evaluation criteria, datasets used, and applications. Table 4 compares the survey with ten earlier surveys. The authors say theirs differs by covering etymology, equations, and a symbolic framework (the nomenclature in Table 1.A).

The main body gives taxonomies in Section 6. They cover deformation type (rigid, affine, deformable), optimization phase, algorithms (Voxelmorph, SynthMorph, ICP, Demons, SyN), correspondence space (atlas-based), correspondence relation, multistage registration (coarse-fine, pyramid), space geometry, and modalities. Section 7 defines evaluation measures in the paper's own notation (Equations 18-28: runtime, MSE, RMSE, TRE, NTRE, Hausdorff distance, centre of mass, Dice). Section 8 (Table 5) lists public datasets with links. Section 9 covers applications: image-guided surgery, tumour diagnosis and

therapy, and motion. Section 10 names transformers and "No Registration" as research directions. Appendix B is a comparison table of surveyed papers with 16 columns, and Appendix C holds ten statistics figures (Figures 29-38).

2. Whether the conclusion holds on the data

What I value most is that the survey can be checked. The search is stated with numbers that add up: $270 - 38 - 41 = 191$ (Section 4). One notation (Table 1.A) runs through Equations 18-28 in Section 7, so TRE, NTRE, Hausdorff distance, and Dice can be read side by side. The 2x2-pixel example in Section 3 lets a newcomer follow a displacement field by hand. Appendix B codes each paper in 16 columns, and Table 5 groups dataset links by region of interest.

On selection, I would add the rule for choosing papers. Section 4 says 96 of 191 papers were reviewed, but I found no rule for choosing the 96. Appendix B, as rendered on Qeios, has 61 rows by my count, so 35 of the 96 ($96 - 61$) do not appear in the table. Could the authors say where they are covered? By my count, the reference list has 110 entries (80 dated 2021 or 2022, 28 before 2021, 2 dated 2023), while Section 4 states 96 papers from 2021-2022 and 10 from before 2021. Here, 80 is less than 96, and 28 is more than 10. The title says deep learning, but in the Approach column of Appendix B, by my count, 34 rows are learning-based, 25 are Classical (22) or Software (3), and 2 are blank. So about 4 in 10 coded papers are not deep learning. The search was limited to open access (41 papers dropped, Section 4) and to one database with one keyword phrase. It would help to discuss the effect of this on coverage. The text also has no limitations section and no conclusion section.

Some tables and text do not agree. Section 9.1 says Ashfaq et al., 2022 aligned 2D inter-operative and 3D preoperative images, but its Appendix B row has Dimensionality "2D". Section 6.7 cites Terpstra et al., 2022 as an example of MIR on 3D point clouds, while its row has Modals "MR" and Dimensionality "2D". Section 10.1 calls Transmorph (Chen, J. et al., 2022) a coarse-fine registration, and its row has Stages "Pyramid". Also, the Appendix C statistics (Figures 29-38) are figures only. The running text never quotes a count or percentage from them, and all percentages in the text belong to other surveys (Section 5).

Several comparative claims have no numbers inside the paper. Section 6.3.1 says Voxelmorph "cut the registration runtime to minutes/seconds compared to hours", and Section 7.1 says "from hours to seconds while keeping almost the same performance". No runtime figure, dataset, or hardware is given, although Section 7.1 itself says a fair comparison needs the same hardware. Section 6.5 gives the Metamorphosis runtime as "10-20 times that of Voxelmorph", with no absolute values. Section 10.1 attributes slow uptake

of transformers to the small number of images in MIR datasets, but Table 5 gives no dataset sizes. A small table of reported values would support these statements.

Finally, the research question. Table 3 asks what evaluation criteria and what datasets were used in 2021-2022. Section 7 defines measures, and Table 5 lists datasets, but Appendix B has no column for evaluation metric or dataset name. So the paper does not show which measures or datasets the surveyed papers actually used. I would add these two columns, and then the question would be answered.

3. Reproducibility

I checked whether another group could repeat the search and rebuild the tables. Much is given: Scopus, the phrase "medical image registration", English, open access, 2021-2022, the counts at each step (270, 38, 41, 191, 96), and the cut-off "as of December 2022" (Section 4). Appendix B shows the coding per paper, so a row can be checked against the source. Table 5 gives a URL for every dataset except IBIS (row "IBIS" has an empty Link cell).

I would add the exact Scopus query string, the date the search was run, and the fields searched, none of which are stated. It would help to list the 270 retrieved, 79 excluded (38 + 41), and 191 remaining records, as no supplementary file is mentioned. There is no rule for choosing 96 from 191, or for choosing the 10 papers from before 2021. Who screened the abstracts and coded the Appendix B columns, and how disagreements were settled, is not said. Definitions of the column values would also help, for example, "Classical" versus "Software" in the Approach column.

Blank cells in Appendix B need explaining. Row "Wang, Z. et al., 2022" is blank from Deformability through Approach, and row "Begum et al., 2022" from Tx function through Stages, so "not reported in the source" and "not coded" look the same. Two rows carry the label "(Upendra, & Hasan et al., 2021)" (ROI Heart, 4D; ROI Blood, 2D), while the reference list has two different Upendra 2021 entries, only one with Hasan as an author. Rows "Wang, D. et al., 2022" and "Zhang, G. et al., 2021" have no entry with those initials in the reference list as rendered on Qeios. Figures 29-38 (Appendix C) are images, so their counts can only be recomputed by re-counting Appendix B. A spreadsheet with the 270 records, the exclusion reason for each, and the Appendix B coding, deposited with a DOI, would make the survey fully checkable.

4. What specifically to fix

1. In Section 4, Appendix B, and the reference list, I would reconcile the numbers: the text reviews 96 papers, Appendix B as rendered on Qeios has 61 rows by my count, and the reference list has 80

entries dated 2021-2022. It would help to state how the 96 were selected from 191. Also, Schmidt & Overhoff is dated 2023 in the reference list (arXiv 2301.08802) but cited as "Schmidt et al., 2022" in Section 9.1 and Appendix B, and a 2023 item falls outside the stated 2021-2022 window.

2. A few cross-references need a fix. The first sentence of Section 5 refers to "the search query in section 6", but the search is described in Section 4. The last subsection of Section 10 is numbered "11.3". Section 6.3.2 (ICP) points to Figure 13 for the ICP demonstration, but the ICP figure is captioned Figure 14, and Figure 13 is the unsupervised approach.
3. As rendered on Qeios, Equation 2 (Section 3.2.1) is followed directly by Equation 6 (Section 6.1), and I could not find Equations 3-5. Were they removed, or lost in rendering?
4. A question on Equation 10 (SyN, Section 6.3.2). The sentence before it writes T_{jk} = inverse of T_{ik} , and Appendix A ("Symmetric Invertibility") agrees with that (T_{ik} = inverse of T_{jk}), but Equation 10 itself, as rendered on Qeios, writes T_{kj} = inverse of T_{ik} . Which subscript order is intended in Equation 10?
5. A question on Equations 11-17 (Section 6.4). As rendered on Qeios, Equation 13 applies T_{ki} to I_j , while the parallel Equation 16 applies T_{kj} to I_k . Equation 17 writes $T_{ki}(I_i) = T_{jk}(I_j)$, where from the definitions I expected $T_{ik}(I_i)$. Are these typos or rendering artefacts?
6. In Section 6.3.1, directly supervised approaches, the text says "The registration field is applied to the fixed image to relocate its pixels". Elsewhere the moving image is the one deformed (Section 3.2.1: "deforming the moving image"; Section 6.6.2 starts from a downscaled moving image). Please check.
7. Table 1.B lists HD as "Hough distance", while Section 7.2 defines HD as "Hausdorff distance". The same table lists nCC, nLCC, SDlogJ, and JOCA, which appear nowhere else in the text. Sections 7.4 (correlation measures) and 7.5 (smoothness) give no numbered equation defining the measure, unlike Sections 7.1-7.3, and the abstract promises correlation-based metrics. I would also shorten the CC scale-invariance remark in Section 7.4, which repeats the wording of Section 6.3.1 almost exactly.
8. For the references, I would align these years between text and list: "Chen, T. et al., (2002)" in Section 6.3.2 vs. 2022; "Nazib et al., 2022" in Section 6.4 vs. 2021 in Appendix B and the list; "Robertson et al., 2022" in Section 9.1 and Appendix B vs. 2021 in the list; "Sutton et al., 1994" in Section 10 vs. 1998. "Wang et al., 2022" without initials (Sections 6.6.2 and 9) is ambiguous because the list has five Wang 2022 entries. As far as I can find, these are listed but never cited in the text or Appendix B: Brown 1992, Bronstein 2021, Goyal 2022, Modersitzki 2003, Oyallon 2015, Mitchell 1980, Ganeshaaraj 2022, Yan 2022, Zhang, J. 2021.

9. In Table 5, NLST and SPARE appear twice (under Chest and under Lung), and IBIS has no link. Adding the number of subjects per dataset, and which surveyed papers used each one, would connect the table to the survey and support the dataset-size argument in Section 10.1.
10. I would add a short Conclusion that answers the question in Table 3 with counts taken from Appendix B, and a Limitations paragraph (open access only, one database, two years). In Section 5, the percentages quoted from Xiao et al. (2021) sum to $43 + 40 + 19 = 102$ for methods and $74 + 18 + 12 = 104$ for metrics, so a note that categories overlap, if they do, would help. A language pass would catch "Synthrmorph" (Section 6.3.1), "daccard" (Section 7.3), "corrosondence" (Equation 21 note), and "wrapped/wrapping" where "warped/warping" seems meant.

5. Recommendation

Major revisions. I would keep the numeric example (Section 3), the single notation across Equations 18–28 (Section 7), and the 16-column Appendix B because they make the tutorial part clear and useful. The survey part needs its numbers reconciled before readers can rely on it. It would help to settle 96 reviewed papers (Section 4) against 61 rows (Appendix B), state the selection rule from 191 to 96, answer the evaluation-criteria and dataset sub-questions of Table 3 from the surveyed papers, and add a conclusion, a limitations section, and a citation and equation-reference cleanup. None of this needs new experiments; the screening list and the coding need to be opened, and the tables recounted.

Declarations

Potential competing interests: No potential competing interests to declare.