C-ARM GUIDANCE: A SELF-SUPERVISED APPROACH TO AUTOMATED POSITIONING DURING STROKE THROMBECTOMY

A. Arrabi[1⋆], *J. Jung*[1⋆], *J. Le*[2], *A. Nguyen*[2], *J. Reed*[2], *E. Stahl*[2], *N. Franssen*[2], *S. Raymond*[3], *S. Wshah*[1]

[1] University of Vermont, Department of Computer Science, Burlington VT, USA
[2]Univeristy of Vermont Medical Center, Burlington VT, USA
[3] Cleveland Clinic, Neurological Institute, Cleveland OH, USA

⋆ Equal contribution

ABSTRACT

Thrombectomy is one of the most effective treatments for ischemic stroke, but it is resource and personnel-intensive. We propose employing deep learning to automate critical aspects of thrombectomy, thereby enhancing efficiency and safety. In this work, we introduce a self-supervised framework that classifies various skeletal landmarks using a regression-based pretext task. Our experiments demonstrate that our model outperforms existing methods in both regression and classification tasks. Notably, our results indicate that the positional pretext task significantly enhances downstream classification performance. Future work will focus on extending this framework toward fully autonomous C-arm control, aiming to optimize trajectories from the pelvis to the head during stroke thrombectomy procedures. All code used is available at https://github.com/AhmadArrabi/C_arm_guidance

***Index Terms*—** Surgical guidance, C-arm positioning, Stroke thrombectomy, Self-supervised, X-ray imaging

1. INTRODUCTION

Many of the most time-sensitive operations for life-threatening conditions such as stroke and trauma require rapid and accurate positioning of fluoroscopy equipment. These procedures are increasingly performed by less experienced operators in low-resource settings [1], creating a need for technologies to improve efficiency and safety.

During stroke treatment, the operator first gains vascular access at the groin and then drives the biplane, two orthogonal X-ray detectors, from a position over the pelvis up to the brain. Fluoroscopy during positioning exposes the patient and operator to extra radiation and takes precious minutes and attention away from other critical tasks. Semi- or automated positioning could reduce the time and radiation required [2], and free the operator to focus on other critical tasks such as device preparation. Thus, we propose the use of deep learning algorithms to assist operators with fluoroscopy operation, particularly in solo- or low-volume settings.

Automated X-ray positioning has been proposed for orthopedic applications [3, 4], focusing primarily on specific body parts, e.g., lumbar spine, proximal femur [5], knee [6], and pelvis [7]. However, its application in stroke treatment remains unexplored due to several challenges: the larger range of motion from the groin to the head, the precise positioning required for cerebral angiographic views, and the broad array of landmarks and radio-dense distractors.

We tackle fluoroscopy positioning in stroke thrombectomy, incorporating a larger anatomical region from the pelvis to the head. Our approach tackles two key tasks, 1) a **classification task** that aids in accurate interpretation of the current biplane position relative to the patient i.e. *what body part are we looking at?*, and 2) a **regression task** that estimates the biplane position based on the X-ray input, i.e., *where are we located?* Our framework presents a step towards broader automation in fluoroscopy control, to ultimately predict the needed trajectory from the current location to the target, i.e. *how do I get to the head from a given location?*

This work presents a self-supervised learning framework to classify 20 skeletal landmarks between the pelvis and the head. Our approach is centered around a positional understanding pretext task, where the model estimates the location of a given X-ray image. Additionally, we embed patient demographic data, that may reveal variations in skeletal structure across patients, into the learned features. Leveraging DeepDRR [8], we developed a custom graphical user interface (GUI) to annotate and generate synthetic X-ray images from computed tomography (CT) scans. We collected two datasets: one includes evenly distributed X-ray images from CTs, for regression, and another, consists of ground truth annotations for classification. All CTs were downloaded from the New Mexico Decedent Image Database (NMDID) [9].

Our contributions can be summarized as follows:

- We work toward automating fluoroscopy positioning (with broader anatomical coverage) for stroke thrombectomy.
- We introduce a self-supervised framework that embeds patient demographic data into its learned features. We find that by leveraging a positional understanding pretext task trained on a large dataset, the performance of the downstream classifier significantly improves.
- We put forward a custom public GUI to simulate C-arm

operation, which can be used as an annotation tool.

2. DATASET

To generate the Digitally Reconstructed Radiographs (DRRs), we used whole-body CT scans from the New Mexico Decedent Image Database (NMDID) [9]. This type of scan covers all regions of interest in stroke thrombectomy, thereby mirroring the continuous movement of the C-arm in procedures. The DRRs were generated by DeepDRR [8], providing synthetic X-ray images.

The NMDID is a public database containing full-body CT scans and demographics of over 30,000 decedents. We selected CT scans with sharper filters (labeled "BONE_") for faster X-ray synthesis with DeepDRR. To ensure full-body coverage, we chose two scan sets, one with arms crossed (labeled "_H-N-UXT_3X3"), and one with arms raised (labeled "_TORSO_3_X_3").

Graphical User Interface (GUI) Annotations As shown in Fig. 1, we developed a custom GUI in Python to facilitate landmark annotation on CT scans. The 3D visualization was implemented using Vedo [10], while the GUI was built with Tkinter [11]. The GUI simulates a C-arm, allowing users to annotate specific landmarks by displaying both 3D views of the CT scan and corresponding DRR images.

Annotated Classification Dataset Utilizing the GUI, we collected annotations of 270 cases (decedents), 47 of which were annotated by radiology residents and the rest were from a non-physician. In total, 8,080 annotations of 20 landmarks were made, Fig. 2 illustrates the location of each landmark and their distribution. Some landmarks are more frequent than others as they exist in both CT scans.

Operators were instructed to click on the position of the assigned landmark. The GUI recorded the C-arm's coordinates through translation only, i.e., 2 degrees of freedom (no rotation, with the z-coordinate fixed). After positioning the C-arm over the landmark, the GUI generated the DRR, which operators could manually fine-tune its position. The following anatomic landmarks were labeled: 1) skull, 2–3) humeral heads, 4–5) scapulas, 6–7) elbows, 8–9) wrists, 10) T1, 11) carina, 12–13) hemidiaphragms, 14) T12, 15) L5, 16–17) iliac crests, 18) pubic symphysis, and 19–20) femoral heads.

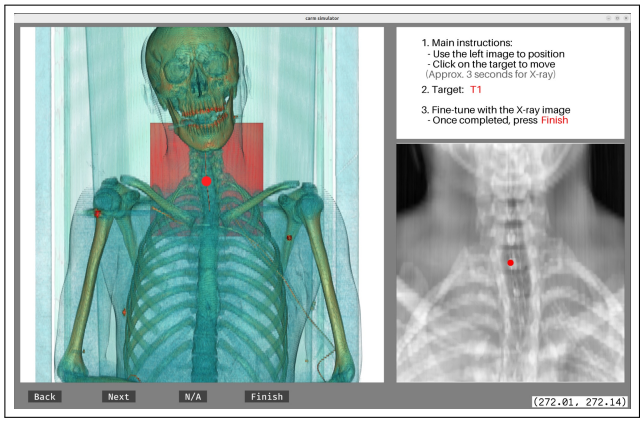

Fig. 1. GUI visualization. The annotators were given direct instructions on collecting the 20 landmarks (top right).

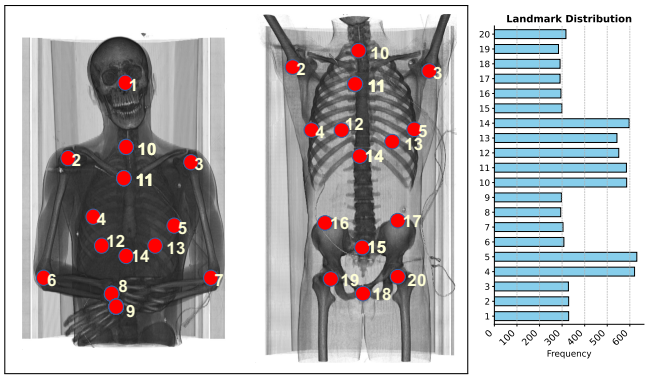

Fig. 2. Landmarks of spatial locations across the skeleton (left), and their distribution in our annotated dataset (right).

Regression Dataset We utilized 392 cases, of which 270 overlap with the ones described above, to generate a larger unannotated dataset for the regression pretext task. We densely sampled CT scans by defining a uniform grid over them, with points spaced 30 mm apart in both vertical and horizontal directions. X-ray images were sampled at these grid points, ensuring systematic and comprehensive coverage of anatomical features. We hypothesize that this approach enables the model to effectively learn the positional relationships of landmarks within the body. A total of 145,338 images were collected, with 76,578 from the one with arms raised CTs and 68,760 from the arms crossed. *As shown in Section 4, the classification and regression tasks were trained on the same cases (decedents), ensuring no overlap between the training and testing sets.*

3. METHODOLOGY

Acquiring large labeled datasets in medical imaging poses significant challenges, e.g., privacy concerns of protected health information, and the need for domain experts for annotations [12]. Self-supervised learning addresses these issues by learning meaningful data representations from large datasets without explicit human labels [13]. A model is first trained on a pretext task to capture the underlying structure of the data and then fine-tuned for the target downstream task, which typically has limited labeled data.

Selecting an appropriate pretext task plays an important role in shaping the learned representations during training. In this work, we propose a regression-based pretext task, where the model predicts the spatial location of a given X-ray image. We hypothesize that this task enables the model to develop a comprehensive understanding of full-body anatomical positioning, particularly of the skeletal structure. Additionally, we propose to embed patient's demographic data into the model, incorporating age, sex, cadaver height, and weight to enrich

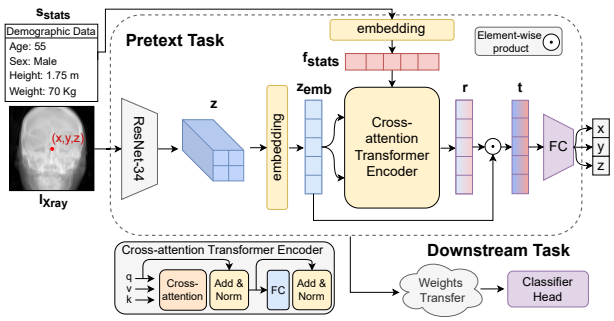

Fig. 3. Main architecture of our proposed framework. The learned representations in the regression task are used to fine-tune the downstream classification task.

the model's learned features.

As illustrated in Eq. (1), we can formalize our method as follows: Given an input X-ray image $I_{X_{ray}}$, located at position $\mathbf{P} = (x, y, z)$ in a predefined coordinate system, and demographic data s_{stats}. Our model g_θ estimates the image location $\mathbf{P}$ given both $I_{X_{ray}}$ and s_{stats}, where θ are the network's learnable parameters. *Note that our chosen origin point $O(0, 0, 0)$ was the top right corner of the CT scan.*

$$g_\theta(I_{X_{ray}}, s_{stats}) = \hat{\mathbf{P}} = (\hat{x}, \hat{y}, \hat{z}) \tag{1}$$

As shown in Fig. 3, $I_{X_{ray}}$ is first projected into a lower-dimensional latent representation $z \in \mathbb{R}^{c \times h \times w}$ using a ResNet-34 [14] backbone, where c, h, and w are the channel, height, and width dimensions, respectively. Then, both s_{stats} and z are projected into continuous vector representations f_{stats}, $z_{emb} \in \mathbb{R}^D$ through fully connected layers, where D is the embedding dimension. We incorporate cross-attention between these two representations, where the query is derived from f_{stats}, and the key and value are obtained from z_{emb}. The output of the cross-attention mechanism is further processed by a standard transformer block [15], producing the vector $r \in \mathbb{R}^D$. This vector is expected to have both spatial and demographic information. To further refine the signal and reduce any noise from f_{stats}, a skip connection from z_{emb} is added, combining it with r via element-wise multiplication to generate the final refined feature $t \in \mathbb{R}^D$. Finally, the regression head consists of a fully connected network with two linear layers that map t to $(\hat{x}, \hat{y}, \hat{z})$. We adopted the Mean Squared Error (MSE) loss to train our model.

After training the regression model, we adapt the architecture for the downstream classification task by replacing the regression head with a classification head. Leveraging the pretrained weights from the pretext task allows the model to preserve rich, domain-specific representations, which enhances its ability to interpret X-ray images. This fine-tuning significantly improves classification performance compared to models initialized with random weights or even pretrained on out-of-domain data, such as ImageNet.

Given that this is a multi-class classification task, we fine-tune the model using the cross-entropy loss. There are several strategies for fine-tuning: one approach involves retraining the entire model, which updates all weights. However, in resource-constrained environments, a more efficient technique known as "linear probing" can be employed. In this method, only the final linear layers are updated and all previous layers are frozen. Linear probing achieves a balance between performance and computational efficiency, as it minimizes the need to retrain the entire model while still adapting to the new task.

4. EXPERIMENTS

Implementation Details: All models were implemented using PyTorch [16]. We used a batch size of 512 trained across 8 AMD Radeon MI50 GPUs in the regression task. For the classification task, the batch size was 64, trained on a single GPU. We employed the Adam optimizer with a learning rate of 0.0001. All X-ray images were resized to 256×256 pixels, and random color jittering and posterization were used for data augmentation. The embedding dimension D for the vectors z_{emb}, r, and t was set to 128.

For evaluation, we benchmark our model on a test set of 54 randomly selected decedents. This split ensures no overlap between the training and testing datasets. For the regression task, we used the MSE between the predicted and the ground truth coordinates, while for classification, we computed micro-averaged precision, recall, and F1-score. Our baseline for both tasks was the modified PoseNet from [5], a method used in previous automated positioning research.

As shown in Table 1, our regression model achieved a mean positional error of 4.7 mm, outperforming PoseNet, reducing the error by approximately 1.2 mm. For classification (Table 2), we retrain the whole network and compared three weight initialization strategies: 1) from the pretext regression task, 2) from ImageNet [17] pretraining, and 3) random initialization. We also included ViT-base [18] and PoseNet as baselines. Our model achieved the best performance across all metrics, with notable improvement when using pretext task weights, highlighting the advantage of domain-specific pretraining over out-of-domain ImageNet weights. Vit-base performed the worst, which we attribute to it requiring more data for optimal performance given it is transformer-based.

Additionally, we explored linear probing by freezing all but the last two linear layers. Table 3 shows a notable performance decline with ImageNet and random weight initializations, demonstrating the benefit of task-specific pretraining.

Finally, we conducted ablation studies to assess the impact of two factors: 1) training fewer layers in linear probing, and 2) removing patient demographics from the model. Table 4 shows that training only one linear layer led to about a 0.2 performance drop, which is expected due to fewer learnable parameters. Furthermore, excluding patient demographic data resulted in a slight performance decline, showing the value of incorporating these features.

Table 1. Regression model performance, best results are bold.

Method	MSE (raw) ↓	MSE (mm) ↓
ours	**0.0084**	**4.7989**
PoseNet [5]	0.0106	6.0717

Table 2. The classification results when retraining all layers of our model. † indicates initializing the weights from the regression task. ‡ indicates ImageNet initialization, and ⋆ is random weight initialization.

Method	Precision ↑	Recall ↑	F1-score ↑
ours†	**0.95**	**0.95**	**0.95**
ours‡	0.94	0.93	0.93
ours⋆	0.93	0.92	0.92
PoseNet [5]	0.91	0.91	0.91
ViT-base [18]	0.79	0.77	0.77

Table 3. The classification results with linear probing on the last two linear layers. A significant gap is present between pretraining on domain-specific and out-of-domain data.

Weight Initialization	Precision ↑	Recall ↑	F1-score ↑
pretext (regression)	**0.83**	**0.83**	**0.82**
ImageNet	0.38	0.41	0.38
Random	0.22	0.20	0.17

Table 4. Ablation studies results. Fine-tuning more linear layers naturally improves performance, and the addition of demographic patient data leads to slight enhancement.

Ablation study	Precision ↑	Recall ↑	F1-score ↑
Two layer linear probing	**0.83**	**0.83**	**0.82**
One layer linear probing	0.62	0.63	0.60
w/ Patient demographics	**0.95**	**0.95**	**0.95**
wo/ Patient demographics	**0.95**	0.94	0.94

5. CONCLUSION AND FUTURE WORKS

We developed a model of human radiographic anatomy using simulated x-rays from a CT database and used the model to classify 20 radiographic landmarks. This lays the groundwork for C-arm positioning during time-sensitive procedures such as stroke thrombectomy. Translation into a clinical setting will require a 3D model that is robust to diverse patient populations and noisy imaging. Future work will expand the model to include rotational and depth information by annotating landmarks on the source CT scans and then training on a large sample of simulated x-rays taken at different C-arm angulation. The model will be further refined and evaluated on clinical fluoroscopy images acquired from biplane positioning during cerebral angiography. Finally, we will evaluate different trajectory strategies to minimize radiation dose and C-arm motion during procedures.

6. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by NM-DID [9]. Ethical approval was not required as confirmed by the license attached with the open access data.

7. ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under Grant No. 2218063.

8. REFERENCES

[1] L. K. Stein, J. Mocco, J. Fifi, N. Jette, S. Tuhrim, and M. S. Dhamoon, "Correlations between physician and hospital stroke thrombectomy volumes and outcomes: A nationwide analysis," *Stroke*, vol. 52, pp. 2858–2865, June 2021. 1

[2] T. De Silva, J. Punnoose, A. Uneri, M. Mahesh, J. Goerres, M. Jacobson, M. D. Ketcha, A. Manbachi, S. Vogt, G. Kleinszig, A. J. Khanna, J.-P. Wolinksy, J. H. Siewerdsen, and G. Osgood, "Virtual fluoroscopy for intraoperative c-arm positioning and radiation dose reduction," *J Med Imaging (Bellingham)*, vol. 5, p. 015005, Feb. 2018. 1

[3] C. Martín Vicario, F. Kordon, F. Denzinger, J. S. El Barbari, M. Privalov, J. Franke, S. Thomas, L. Kausch, A. Maier, and H. Kunze, "Automatic plane adjustment of orthopedic intraoperative flat panel detector CT-volumes," *Journal of Medical Imaging (Bellingham, Wash.)*, p. 034001, May 2022. 1

[4] L. Kausch, S. Thomas, H. Kunze, T. Norajitra, A. Klein, L. Ayala, J. El Barbari, E. Mandelka, M. Privalov, S. Vetter, A. Mahnken, L. Maier-Hein, and K. Maier-Hein, "C-arm positioning for standard projections during spinal implant placement," *Medical Image Analysis*, p. 102557, Oct. 2022. 1

[5] L. Kausch, S. Thomas, H. Kunze, M. Privalov, S. Vetter, J. Franke, A. H. Mahnken, L. Maier-Hein, and K. Maier-Hein, "Toward automatic C-arm positioning for standard projections in orthopedic surgery," *International Journal of Computer Assisted Radiology and Surgery*, no. 7, pp. 1095–1105, 2020. 1, 3, 4

[6] L. Kausch, S. Thomas, H. Kunze, J. S. El Barbari, and K. H. Maier-Hein, "Shape-based pose estimation for automatic standard views of the knee," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023* (H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor, eds.), (Cham), pp. 476–486, Springer Nature Switzerland, 2023. 1

[7] H. Esfandiari, S. Andreß, M. Herold, W. Böcker, S. Weidert, and A. J. Hodgson, "A deep learning approach for single shot c-arm pose estimation," in *CAOS 2020. The 20th Annual Meeting of the International Society for Computer Assisted Orthopaedic Surgery* (F. R. Y. Baena and F. Tatti, eds.), EPiC Series in Health Sciences, pp. 69–73, EasyChair, 2020. 1

[8] M. Unberath, J.-N. Zaech, S. C. Lee, B. Bier, J. Fotouhi, M. Armand, and N. Navab, "Deepdrr – a catalyst for machine learning in fluoroscopy-guided procedures," 2018. 1, 2

[9] H. J. Edgar, S. Daneshvari Berry, E. Moes, N. Adolphi, P. Bridges, and K. Nolte, "New mexico decedent image database," 2020. 1, 2, 4

[10] M. Musy *et al.*, "vedo, a python module for scientific analysis and visualization of 3D objects and point clouds," 2019. 2

[11] F. Lundh, "An introduction to tkinter," 1999. http://www.pythonware.com/library/tkinter/introduction/index.html. 2

[12] S. Shurrab and R. Duwairi, "Self-supervised learning methods and applications in medical imaging analysis: A survey," *PeerJ Computer Science*, p. e1045, 2022. 2

[13] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 4037–4058, 2021. 2

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 3

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 6000–6010, Curran Associates Inc., 2017. 3

[16] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017. 3

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009. 3

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. 3, 4