A POINT IN THE RIGHT DIRECTION: VECTOR PREDICTION FOR SPATIALLY-AWARE SELF-SUPERVISED VOLUMETRIC REPRESENTATION LEARNING

Yejia Zhang† * Pengfei Gu† * Nishchal Sapkota† Hao Zheng† Peixian Liang† Danny Z. Chen†

†University of Notre Dame, Department of Computer Science and Engineering, Notre Dame, IN, USA

ABSTRACT

High annotation costs and limited labels for dense 3D medical imaging tasks have recently motivated an assortment of 3D self-supervised pretraining methods that improve transfer learning performance. However, these methods commonly lack spatial awareness despite its centrality in enabling effective 3D image analysis. More specifically, position, scale, and orientation are not only informative but also automatically available when generating image patches for training. Yet, to date, no work has proposed a pretext task that distills all key spatial features. To fulfill this need, we develop a new self-supervised method, VectorPOSE, which promotes better spatial understanding with two novel pretext tasks: Vector Prediction (VP) and Boundary-Focused Reconstruction (BFR). VP focuses on global spatial concepts (i.e., properties of 3D patches) while BFR addresses weaknesses of recent reconstruction methods to learn more effective local representations. We evaluate VectorPOSE on three 3D medical image segmentation tasks, showing that it often outperforms state-of-the-art methods, especially in limited annotation settings.

Index Terms— 3D Self-Supervised Learning, Representation Learning, Volumetric Medical Image Segmentation.

1. INTRODUCTION

Modern 3D medical image analysis techniques have made great performance strides by extracting appearance-based features from local image patches. However, they fail to explicitly attend to a key capability that facilitates volumetric analysis: spatial awareness. Given an arbitrary image patch (e.g., see Fig. 1), we can often leverage anatomical priors and infer a large amount of information from its position, scale, and orientation. Patch position such as its axial window (e.g., the dashed red and blue lines in Fig. 1 that span the CT volume’s depth) reveals which anatomical structures to expect. Image zoom is commonly adjusted by practitioners, and the knowledge of a patch’s scale is essential for assessments of structure size, shape, and extent. Finally, patch orientation informs object poses and valid adjacent structures. From these insights, we hypothesize that networks which better predict 3D patch position, scale, and orientation can more accurately segment images with superior anatomical knowledge.

* Equal Contributions.
To remedy this gap, we propose a new 3D self-supervised pretraining approach, VectorPOSE (POSE for Position, Orientation, Scale, and Extent), with two novel pretext tasks for learning improved spatial and appearance features. Our first pretext task, Vector Prediction (VP), infers vectors that originate from predefined points within a sampled input patch (e.g., patch center, corners) and terminate at the center of the volume. Vectors encapsulate both magnitude and direction, so predicting multiple vectors with consistent origin points effectively encodes position (with respect to the terminal point), orientation (via vector angles to the terminal point), and scale (where differentials between predicted vector origins represent patch extent). To learn effective local features and address boundary degradations in reconstruction, our second pretext task, Boundary-Focused Reconstruction (BFR), conducts: 1) boundary reconstruction where extracted edges of the patch are explicitly predicted, and 2) voxel reconstruction where original intensity values are predicted with a loss that better preserves boundaries. To demonstrate efficacy, we pretrain and evaluate on three 3D semantic medical image segmentation tasks (i.e., CT cardiac, CT abdominal, MR prostate). VectorPOSE generally outperforms state-of-the-art single- and multi-task methods while exhibiting superior data efficiency compared to known contrastive learning methods. Our contributions are summarized below.

1. We propose a self-supervised approach, VectorPOSE, that places a holistic emphasis on three key spatial properties (position, orientation, scale) while synergizing learned global information with local, boundary-aware cues.

2. As a cornerstone to spatial understanding, we introduce a novel Vector Prediction (VP) pretext task that succinctly encapsulates key spatial properties and is both effective and extensible. Further, we propose the Boundary-Focused Reconstruction (BFR) pretext task to address pitfalls of recent works and learn improved local features.

3. VectorPOSE is comprehensively evaluated on two 3D CT segmentation datasets (MMWHS [15, 16] & BCV [17]) and one 3D MR prostate dataset [18]. Not only does it outperform recent self-supervised methods in the majority of settings, it also exhibits improved data efficiency without additional training time.

2. METHODOLOGY

Before detailing the proposed pretext tasks, we first overview our general pipeline and its major components. Following established self-supervised methods, we pretrain a randomly initialized neural network consisting of an encoder E, a decoder D, the prediction head $D_{vp}$ for Vector Prediction (VP), and the prediction head $D_{bfr}$ for Boundary-Focused Reconstruction (BFR). Note that $D_{vp}$ employs global information and processes pooled & flattened output features from E, while $D_{bfr}$ utilizes D’s output which has the same resolution as E’s input. To pretrain $\{E, D, D_{vp}, D_{bfr}\}$ with VectorPOSE (see Fig. 2), we input patches $x^{(i)}$, $x^{(j)}$, … from the ith 3D image $X^{i}$ of an unlabeled dataset $D_{u}$. Stochastic spatial augmentations $T_{S}$ (e.g., flipping or rotating along each axis) and intensity noises $T_{I}$ (e.g., local pixel shuffle, masking, intensity shifting) are applied to all input patches independently and in order ($T_{S}$ first and then $T_{I}$). The targets for VP & BFR are computed after $T_{S}$ is applied. After pretraining, the prediction heads $D_{vp}, D_{bfr}$ are discarded while $E, D$ are transferred and fine-tuned on a downstream target task.

2.1. Vector Prediction (VP)

The objective of VP is to distill position, scale, and orientation information from patches by predicting multiple vectors which all point to one standardized location existing in all dataset volumes (e.g., the volume center or a keypoint from a chosen structure). We refer to this point as the reference landmark, $(x^{(i)}, y^{(i)}, z^{(i)})$, which denotes the landmark coordinate for patch $x^{(i)}$ from 3D image $X^{i}$. Reference landmarks across images have anatomical correspondences from which we learn useful priors and structural consistencies. In many 3D tasks, reference landmarks can simply be taken as the
center of volumes given rough image alignments. In others, positions of landmarks can be extracted in an unsupervised manner (e.g., center of the left lung as a landmark after lung masks are automatically extracted [19]). To prevent overfitting to this single point, we add regularization and spatially jitter the reference landmark coordinates with random perturbations within $\eta\%$ of each image dimension’s total length.

After the landmark is extracted for every volume, we select each patch’s origin points. A vector’s origin point concretely defines distance (via its magnitude) and position (via angles) to the reference. For a sampled patch $x^{(ij)}$, we define $n$ vector origin points as $(x_{0}^{(ij)}, y_{0}^{(ij)}, z_{0}^{(ij)})$, $(x_{1}^{(ij)}, y_{1}^{(ij)}, z_{1}^{(ij)})$, ..., $(x_{n-1}^{(ij)}, y_{n-1}^{(ij)}, z_{n-1}^{(ij)})$. In practice, we set these origin points as consistent positions in all patches (e.g., first origin point at the patch center, second at the patch’s top-right corner, etc.). Mathematically, predicting a single vector is sufficient to describe position and orientation, while two are enough to also cover scale. However, we empirically find that $n=9$ vectors (vectors originating from the patch center and 8 corners) improve feature learning (see Section §3, Tab. 2).

Thus, all vectors belonging to patch $x^{(ij)}$ are defined as $v^{(ij)}_m$ ($m=0, \ldots, n-1$) with origin point $(x_m^{(ij)}, y_m^{(ij)}, z_m^{(ij)})$ and terminal point $(x^m, y^m, z^m)$. The terminal points are identical for all patches from the same image. Below we omit “$m$” for readability. For prediction, we reparameterize vectors into spherical coordinates: let $\phi_m=\arctan(y_m/x_m)$, $\theta_m=\arccos(z_m/r_m)$, and $r_m=\sqrt{x_m^2+y_m^2+z_m^2}$.

The overall VectorPOSE loss for an image patch is:

$$L_{vp} = \frac{1}{n} \sum_{m=0}^{n-1} \left( \| \hat{r}_m - \sigma(\hat{r}_m) \| + \| \hat{\theta}_m - \sigma(\hat{\theta}_m) \| + \min(\| \phi_m - \tanh(\phi_m) \|, \| \phi_m - \tanh(\phi_m) + 2\pi \|) \right)$$

(1)

where $\sigma$ is the sigmoid function, $R$ is the radius of a sphere circumscribing the image volume, and $\hat{r}_m, \hat{\theta}_m, \phi_m$ are logits from $E \circ D \circ D_{vp}$ (see the bottom right of Fig. 2). To address the $\phi$ angles that are physically close but distant in the parameter space (e.g., $-179^\circ$ & $+179^\circ$), we take the minimum among targets $\phi_m, \phi_m - 2\pi, \phi_m + 2\pi$ (see Eq. (1)). Also, note that when patches undergo spatial augmentations (e.g., flipping, rotations), the index $m$ of each vector’s post-transform origin corresponds with the origin before transforming (see “Vector Creation” in Figure 2).

By encapsulating the three key spatial properties into an interdependent task, we prevent training instability and feature interference from independent categorical predictions [10]. Additionally, vectors represent spatial properties in a continuous space and is a generalized extension of previous tasks like discrete scale & rotation prediction [13].

### 2.2. Boundary-Focused Reconstruction (BFR)

Two appearance-focused losses are employed in BFR. To address boundary degradation and reduced texture acuity, we first switch the reconstruction criterion from $L_2$ loss (as used in [9]) to $L_1$ loss, which has been shown to improve both reconstruction quality and boundary accuracy [11]. To further emphasize anatomical delineation, we add an additional boundary reconstruction task where boundaries are extracted via a 3D Scharr edge detector. We select the Scharr transform since it exhibits superior rotational invariance compared to other filters like Sobel. In conjunction with strong augmentations such as texture noising (e.g., local pixel shuffle) and masking (e.g., in-painting or out-painting), explicit edge reconstruction facilitates the learning of anatomical shapes and boundary extents. The overall BFR loss is defined as:

$$L_{bfr} = \frac{1}{n} \sum_{m=0}^{n-1} \left( \| y^\alpha_m - \sigma(y^\alpha_m) \| + \| y^b_m - \sigma(y^b_m) \| \right),$$

(2)

where $y^\alpha_m$ and $y^b_m$ are the targets of the voxel and boundary reconstruction tasks, $\hat{y}^\alpha_m$ and $\hat{y}^b_m$ (from $E \circ D \circ D_{bfr}$) are the voxel and boundary logits, respectively. $\alpha$ is a scaling term to address the small proportion of boundaries in patches.

### 2.3. Overall Loss and Implementation

The overall VectorPOSE loss for an image patch is:

$$\mathcal{L} = \lambda L_{bfr} + (1 - \lambda) L_{vp},$$

(3)

For the UNet-like model, we employ a 3D version of ResNet-50 [20] as the encoder and attach it to a light decoder with trilinear upsampling layers & additions for feature fusion. For $D_{vp}$, the lowest encoder output is pooled, flattened, and processed through a 2-layer MLP with 256 hidden dimensions, while $D_{bfr}$ uses a single 1x1x1 convolution. We use $m=9$ vectors per patch with jitter $\eta=5\%$, and train with $\alpha=5$ & $\lambda=0.5$.

### 3. EXPERIMENTS & RESULTS

We implemented all experiments using PyTorch with NVIDIA Tesla P100s. For pretraining, volume intensities were normalized between 0 and 1, uninformative regions were excluded, (96, 96, 96) patches were sampled, and augmentations in [9] were applied. We trained using AdamW (0.0002 learning rate, 0.0001 weight decay) for 300 epochs (~21 hours) with a batch size of 12. For fine-tuning, we normalized volume intensities between 0 and 1, sampled (64, 128, 256) patches, and applied random flipping, brightness, gamma, and blurring. We optimized using AdamW (0.001 learning rate, 0.001 weight decay) for 200 epochs (~14 hours) with a batch size of 4. For evaluation, we reported the average F1 scores across foreground classes over 8 runs. Inference utilized sliding window (30% overlap) with the same patch size as training.

#### 3.1. Datasets

All the datasets were split using a 6:2:2 training:validation:test ratio. Models were pretrained using the full training sets (no
Table 1. Class-averaged Dice score (%) comparisons vs. recent self-supervised 3D medical image segmentation methods. Entries that are bolded & underlined are the best & second-best scores, respectively.

| Method                | CT MMWHS | CT BCV | MRI MSD-Prostate |
|-----------------------|----------|--------|------------------|
|                       | 10%  | 25%  | 50%  | 100%          | 10%  | 25%  | 50%  | 100%          | 10%  | 25%  | 50%  | 100%          |
| Random Init.          | 79.31 | 87.29 | 89.15 | 90.70          | 30.06 | 58.48 | 66.23 | 75.81          | 40.05 | 58.78 | 69.25 | 74.24          |
| Models Genesis [9]  | 80.05 | 88.14 | 90.54 | 91.05          | 32.03 | 58.60 | 66.87 | 75.79          | 44.00 | 60.89 | 69.55 | 75.86          |
| SAR [10]              | 81.48 | 88.57 | 90.81 | 91.14          | 35.63 | 60.29 | 67.52 | **76.77**      | 42.53 | 61.02 | 70.16 | 74.62          |
| Rubik++ [11]          | 81.08 | 88.40 | 90.94 | 91.23          | 33.28 | 59.83 | 67.72 | 76.65          | 43.70 | 61.97 | **71.35** | 75.81          |
| PGL [21]              | 79.76 | 87.50 | 89.48 | 90.57          | 30.99 | 58.56 | 66.39 | 75.98          | 41.36 | 59.08 | 69.97 | 74.08          |
| MoCo [4]              | 80.22 | 87.94 | 89.60 | 91.05          | 31.89 | 58.95 | 66.91 | 76.01          | 42.07 | 60.69 | 70.17 | 74.31          |
| VectorPOSE (Ours)     | **83.30** | **89.43** | **91.27** | **91.60** | **37.88** | **61.61** | **68.01** | **76.53** | **46.35** | **63.14** | **71.22** | **75.97** |

Table 2. Ablations on the proposed components on 10% MMWHS. The top two rows contain appearance-based components while the middle two rows describe VP settings.

| Method   | CT MMWHS | CT BCV | MRI MSD-Prostate |
|----------|----------|--------|------------------|
| Voxel Rec. | $L_1$ | $L_1$ | $L_1$ | $L_1$ | $L_1$ | $L_1$ | $L_1$ | $L_1$ | $L_1$ | $L_1$ | $L_1$ | $L_1$ | $L_1$ |
| Bound. Rec. | $L_1$ | $L_1$ | $L_1$ | $L_1$ | $L_1$ | $L_1$ | $L_1$ | $L_1$ | $L_1$ | $L_1$ | $L_1$ | $L_1$ | $L_1$ | $L_1$ |
| Center Vector | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Corner Vector(s) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Dice (%) | 80.16 | 80.80 | 81.37 | 82.29 | 82.75 | **83.30** |

3.2. Results and Discussion

For fair comparisons, all methods were pretrained and fine-tuned using 10%, 25%, 50%, or 100% of the training data.

CT MMWHS [15, 16] provides 20 annotated CT volumes segmenting seven cardiac structures (left ventricle, left atrium, right ventricle, right atrium, myocardium, ascending aorta, and pulmonary artery). CT BCV [17] is an abdominal organ segmentation challenge with 30 annotated CT volumes and 13 classes. MRI MSD-Prostate [18] contains 32 labeled MRI (T2, ADC) volumes with two foreground classes.

3.3. Ablations

Here, we explore the contributions of each individual component in our method. Notably, we intend to test our hypothesis about the importance of explicitly learning spatial parameters and compare its efficacy against other proposed methods (e.g., reconstruction). In Table 2, one can see ~0.6% improvement after adding the boundary awareness task, which supports the benefit of explicitly predicting boundaries over voxel reconstruction alone. Next, we see that with increasing $m$ (the number of vectors per patch), performance notably rises. This supports our hypothesis of using spatial awareness as an effective feature-learning principle.

4. CONCLUSIONS

We proposed VectorPOSE, a new self-supervised approach for 3D medical image segmentation with two novel pretext tasks: 1) Vector Prediction (VP) to address the lack of spatial understanding (i.e., position, scale, orientation) in existing works; 2) Boundary-Focused Reconstruction (BFR) for improving the understanding of anatomical structure extents through boundary regression. After evaluation on three 3D semantic segmentation datasets, VectorPOSE's general outperformance over state-of-the-art methods shows that spatial understanding is pertinent for feature learning and may be a promising avenue for future research progress.
5. COMPLIANCE WITH ETHICAL STANDARDS

This research was conducted retrospectively using open-access human subject data by three publicly available datasets [16–18]. Ethical approval was not required as confirmed by the licenses attached to the open-access datasets.

6. REFERENCES

[1] A. Taleb, W. Loetzsch, N. Danz, J. Severin, T. Gaertner, B. Bergner, and C. Lippert, “3D self-supervised methods for medical imaging,” NIPS, vol. 33, pp. 18158–18172, 2020.

[2] H. Zheng, J. Han, H. Wang, L. Yang, Z. Zhao, C. Wang, and D. Chen, “Hierarchical self-supervised learning for medical image segmentation based on multi-domain data aggregation,” in MICCAI, 2021.

[3] D. Zeng, Y. Wu, X. Hu, X. Xu, H. Yuan, M. Huang, J. Zhuang, J. Hu, and Y. Shi, “Positional contrastive learning for volumetric medical image segmentation,” in MICCAI, 2021.

[4] X. Chen, L. Yao, T. Zhou, J. Dong, and Y. Zhang, “Momentum contrastive learning for few-shot COVID-19 diagnosis from chest CT images,” Pattern Recognition, vol. 113, pp. 107826, 2021.

[5] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, “Contrastive learning of global and local features for medical image segmentation with limited annotations,” NIPS, vol. 33, 2020.

[6] H. Zhou, C. Lu, S. Yang, X. Han, and Y. Yu, “Preservation learning improves self-supervised medical image models by reconstructing diverse contexts,” ICCV, pp. 3479–3489, 2021.

[7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in ICML. PMLR, 2020, pp. 1597–1607.

[8] F. Matzkin, V. F. J. Newcombe, S. Stevenson, A. Khetani, T. Newman, R. Digby, A. Stevens, B. Glocker, and E. Ferrante, “Self-supervised skull reconstruction in brain CT images with decompressive craniectomy,” ArXiv, vol. abs/2007.03817, 2020.

[9] Z. Zhou, V. Sohda, M. Siddiquee, R. Feng, N. Tajbakhsh, M. Gotway, and J. Liang, “Modelsogenesis: Generic autodidactic models for 3D medical image analysis,” MICCAI, vol. 11767, pp. 384–393, 2019.

[10] X. Zhang, S. Feng, Y. Zhou, Y. Zhang, and Y. Wang, “SAR: Scale-aware restoration learning for 3D tumor segmentation,” in MICCAI, 2021.

[11] X. Tao, Y. Li, W. Zhou, K. Ma, and Y. Zheng, “Revisiting Rubik’s cube: Self-supervised learning with volume-wise transformation for 3D medical image segmentation,” in MICCAI. Springer, 2020, pp. 238–248.

[12] X. Zhuang, Y. Li, Y. Hu, K. Ma, Y. Yang, and Y. Zheng, “Self-supervised feature learning for 3D medical images by playing a Rubik’s cube,” in MICCAI. Springer, 2019, pp. 420–428.

[13] J. Zhu, Y. Li, Y. Hu, K. Ma, S. Zhou, and Y. Zheng, “Rubik’s Cube+: A self-supervised feature learning framework for 3D medical image analysis,” Medical Image Analysis, vol. 64, pp. 101746, 2020.

[14] F. Haghhighi, M. Taher, Z. Zhou, M. Gotway, and J. Liang, “Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning,” IEEE TMI, vol. 40, pp. 2857–2868, 2021.

[15] X. Zhuang and J. Shen, “Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI,” Medical Image Analysis, vol. 31, pp. 77–87, 2016.

[16] X. Zhuang, “Challenges and methodologies of fully automatic whole heart segmentation: A review,” Journal of Healthcare Engineering, vol. 4, pp. 371–407, 2013.

[17] “Multi-atlas labeling beyond the cranial vault - workshop and challenge,” 2015, Accessed January 2021 at https://www.synapse.org/#!Synapse:syn3193805/wiki/89480.

[18] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Ginneken, A. Kopp-Schneider, B. A. Landman, G. J. S. Litjens, B. H. Menze, O. Ronneberger, R. M. Summers, P. Bilic, P. F. Christ, R. K. G. Do, M. J. Gollub, J. Golia-Pernicka, S. Heckers, W. R. Jarnagin, M. McHugo, S. Napel, E. Vorontsov, L. Maier-Hein, and M. J. Cardoso, “A large annotated medical image dataset for the development and evaluation of segmentation algorithms,” ArXiv, vol. abs/1902.09063, 2019.

[19] B. Zhao, G. Gamsu, M. S. Ginsberg, L. Jiang, and L. H. Schwartz, “Automatic detection of small lung nodules on CT utilizing a local density maximum algorithm,” Journal of Applied Clinical Medical Physics, vol. 4, no. 3, pp. 248–260, 2003.

[20] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” CVPR, pp. 770–778, 2016.

[21] Y. Xie, J. Zhang, Z. Liao, Y. Xia, and C. Shen, “PGL: Prior-guided local self-supervised learning for 3D medical image segmentation,” ArXiv abs/2011.12640, 2020.