How to learn from unlabeled volume data: Self-Supervised 3D Context Feature Learning

Maximilian Blendowski[1], Hannes Nickisch[2], and Mattias P. Heinrich[1]

[1] Institute of Medical Informatics, University of Lübeck, Germany
{blendowski,heinrich}@imi.uni-luebeck.de
[2] Philips Research Hamburg, Germany

Abstract. The vast majority of 3D medical images lacks detailed image-based expert annotations. The ongoing advances of deep convolutional neural networks clearly demonstrate the benefit of supervised learning to successfully extract relevant anatomical information and aid image-based analysis and interventions, but it heavily relies on labeled data. Self-supervised learning, that requires no expert labels, provides an appealing way to discover data-inherent patterns and leverage anatomical information freely available from medical images themselves. In this work, we propose a new approach to train effective convolutional feature extractors based on a new concept of image-intrinsic spatial offset relations with an auxiliary heatmap regression loss. The learned features successfully capture semantic, anatomical information and enable state-of-the-art accuracy for a k-NN based one-shot segmentation task without any subsequent fine-tuning.

Keywords: Self-Supervised Learning · Volumetric Image Segmentation

1 Introduction and Related Work

Deep learning with convolutional networks (DCNN) has become a powerful and versatile tool for a large variety of medical image analysis tasks. DCNNs stand out with their ability to learn informative features that are robust to artifacts or noise and which do not rely on hand-crafted feature engineering and explicit domain knowledge. However, up to date, nearly all deep networks require large datasets with strong supervision through expert annotations. In contrast to computer vision, tasks where layman can cost-effectively label abundantly available images at low cost are rare in medical imaging [7].

Thus, a large fully-annotated high-quality training corpus is rarely available in medical imaging, which triggered research to relax this assumption in various ways. Weak labels can enable registration tasks [4], noisy labels allow for classification tasks [9], a few labels suffice for segmentation tasks [10] and transfer learning on data from a different domain [11] can be used to detect lung nodules.

In the quest to – ultimately – use unlabeled data for learning, the computer vision community recently explored *self-supervision*, a form of unsupervised learning, where auxiliary tasks are derived from unlabeled data enabling a

machine to extract visual knowledge. These auxiliary tasks are usually easy to verify but require a certain degree of image understanding. Prominent examples are hole filling (inpainting), the prediction of spatial neighborhood relations of image patches [2], the colorization of grayscale images [15] or a combination of a number of them [3].

Applications of self-supervision to medical imaging range from leveraging follow up scans in spine MRI [6] over surrogate supervision used for segmentation from only a fraction of the labels [12] to unsupervised learning employed for image registration [14]. Our work is closely related to the context prediction of neighbouring patches introduced by Doersch et al. [2], which demonstrated the capabilities of using spatial relations (e.g. top/bottom, left/right) that already are inherently given as auxiliary task to pretrain feature extractors in natural, two-dimensional images. The large variety of details and presence of multiple relational objects in natural 2D images enabled them to learn CNNs that extract semantically meaningful descriptors. To ensure a sufficiently demanding self-supervision task, the image patches cannot have any overlap and must also contain recognizable object parts. When considering volumetric medical scans

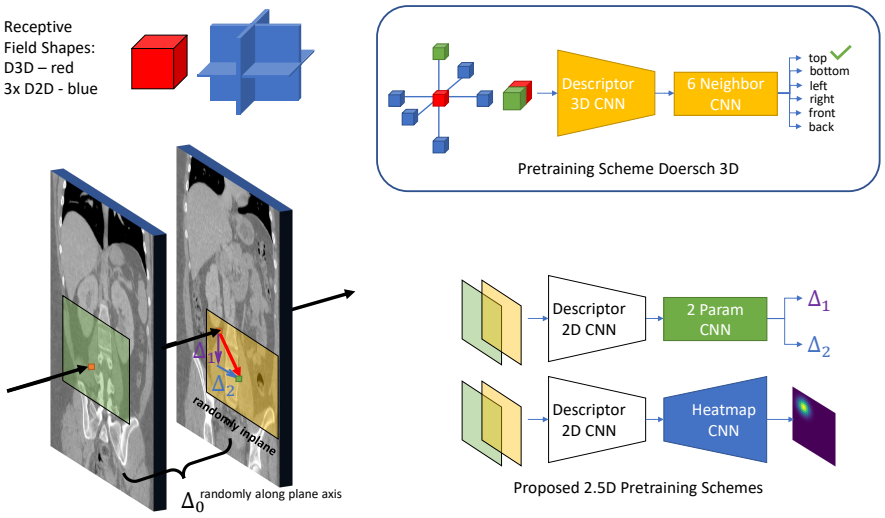


Fig. 1. Top left: inspired by [2] and extended to 3D, DOERSCH uses a cubic receptive field, while our proposed method uses three intersecting planar volumes (2.5D). Top right: DOERSCH ■ predicts the spatial arrangement of two cubic image subvolumes inside a 6 neighborhood as auxiliary task to pretrain *Descriptor 3D CNN*. Bottom left - our approach: per axis predict small continuous-valued offsets (Δ_1, Δ_2) between the centers of disjoint planar volumes that are Δ_0 apart in order to pretrain *Descriptor 2D CNN*. Bottom right: Different ways to implement the auxiliary offset prediction task. 1) REG2D ■: Direct regression of both parameters with fully connected layers in *2 Param CNN*. 2) HEATMAP ■: Regressing (Δ_1, Δ_2)-heatmaps using transposed convolutions in *Heatmap CNN*.

(CT, MRI) there exists a conflicting relation between an increasing size of the patch for the CNN (equivalently its receptive field), which is necessary to capture enough spatial information for an expressive feature learning and a suitable difficulty of the auxiliary task: i.e. the learning task can become too easy, when the receptive field grows, because neighboring subvolumes are likely to contain easily identifiable structures, e.g. body borders ('body border problem').

As illustrated in the upper part of Fig. 1, the *Doersch*-inspired pre-training scheme randomly extracts two 3D subvolumes (red box and green box) from within a six neighbourhood of a considered scan. Both intensity patches are fed into a Siamese convolutional network that yields one feature vector each. These are concatenated and used within a conventional fully-connected network to predict the spatial relation of the two patches as a categorical six-class task.

2 Methods

We strongly believe that a simple extension of a spatial patch-based context prediction to 3D does not fully exploit the potential of self-supervised pre-training for medical scans. Consequently, we introduce a novel method that is inspired by the work of Doersch et al. [2], but aims to overcome the trade-off between the receptive field limitations imposed by unsuitable pretext problems.

Contributions: 1) We propose a new scheme to appropriately leverage spatial information in 3D scans by predicting orthogonal offsets of two large planar patches that are extracted with a small intermediate gap and enables the use of more flexible auxiliary tasks. 2) We use an auxiliary decoder network for 2D heatmap regression that increases the robustness of this offset computation.

2.1 Self-supervised feature learning

The lower part of Fig. 1 illustrates the basic ideas of our work that introduces a new unsupervised pre-training scheme based on spatial cues. Instead of relying on cubical patches (as done in [2]), we propose to extract two nearly planar 2.5D subvolumes along the main imaging axis (e.g. the coronal plane in Fig. 1, approx. 117x97x9mm) with a fixed spatial offset of Δ_0, chosen large enough so that no overlap exists. The anchor patch (green box) is extracted around the voxel of interest in the first slice, while the second patch (yellow box) is randomly shifted in its position along the normal (inplane) direction with continuously drawn offsets (Δ_1, Δ_2) (purple and blue). Since the second slice shares *no* obvious spatial hints (e.g. continuing lines) with the anchor patch, we are no longer limited to few discrete neighbourhood relations and can consider a greater variability of displaced patches. In essence, compared to [2], shrinking the cubical patches along one dimension allows us to avoid the 'body border problem': due to the unknown offset Δ_0, we are able to present diverse image pairs to our networks that provide a sufficiently large receptive field (2D, perpendicular to the Δ_0 offset axis) to inherently learn anatomical information. As before a Siamese convolutional architecture (denoted as *Descriptor 2D CNN* (D2D-CNN)) is trained to

extract vector-valued descriptors for both patches individually. While the cross entropy (CE-loss) for a six-class prediction was a natural choice as loss function for the *Doersch*-inspired 3D pre-training method, we propose to formulate our continuous offset approach as the following auxiliary learning task: Predict the two offset parameters (Δ_1, Δ_2) with heatmap regression (seen to provide a more informative gradient flow in [8]) using an expanding decoder network with transposed convolutions.

The proposed planar patch offset prediction is not limited to a certain axis of a 3D volume and hence three separate D2D-CNN networks are trained in parallel. The final descriptor for a voxel positioned at the intersection of each of the three planar subvolumes is obtained by simply concatenating the output of all three D2D-CNNs. For the sake of a clear notation, we assume all image axes to be normalized to $[-1, 1]$, i.e. with a side length of 2 in the following.

Details on Heatmap ■ network training[3]: We train our proposed 2.5D feature extractor D2D-CNNs (1 per axis) paired with one *Heatmap CNN* each. Following the scheme visually presented in the lower part of Fig. 1, we sample a near planar subvolume represented as a 3-channel 2D image for each axis. These slices have dimensions of 3×42^2 with side lengths 0.8, a depth of 0.05 in the normal direction and form the input of the feature CNN. The anchor slice's central position within the scan is uniformly drawn from $[-0.5, 0.5]^3$. The second subvolume is sampled so that it is displaced by at least $\Delta_0 = 0.125$ and up to $\Delta_0 = 0.25$ in normal direction. This perpendicular offset is *not used* during the training process. The inplane offset parameters (Δ_1, Δ_2) that are the target of the auxiliary learning task are uniformly drawn from $\pm[0.25, 0.3]^2$ in the beginning and up to $\pm[0, 0.7]^2$ at the end of the training process. Enforcing offsets of at least ± 0.25 initially accelerates the context learning. The MSE-loss between the network's predicted heatmaps and the ground truth was used as a penalty term. The heatmap is obtained from the offsets using

$$heat_{gt}(i, j, \Delta_1, \Delta_2) = 10 \cdot e^{-15 \cdot \left[(i/9 - \Delta_1)^2 + (j/9 - \Delta_2)^2\right]}$$

with $(i, j) \in \{-9, -8, ..., +8, +9\}^2$, yielding 19×19 sized images. The final 2.5D descriptors for both methods result from the concatenation of all 3 D2D-CNNs (axial, coronal and sagittal axes).

Implementation of the comparitive 3D Doersch ■ approach: We combine a 3D convolutional network (D3D-CNN) as feature extractor with a six-class prediction network *6 Neighbor CNN* as a straight-forward 3D extension of [2]. The 6 possible neighbouring relations define the auxiliary task trained with a CE-loss. We extract an anchor 3D subvolume of 25^3 voxels as cubes with sidelength 0.4 - its center is again uniformly sampled in $[-0.5, 0.5]^3$ within the image volume in order to be positioned inside the patient's body. As partner, we randomly sample one of its 6 neighboring subvolumes and add jitter to its center coordinates to avoid e.g. line continuation hints [2].

[3] We will release code and pre-trained networks as well as detailed data preprocessing steps to enable reproducibility.

CNN	D2D	2 Param	Heatmap	D3D	6 Neighbor
Input	image data	D2D features	D2D features	image data	D3D features
Layer 1	Conv(3,32,3,1) MP(2,2),GN,LR	Conv(128,128,1,1) GN,LR	Conv(128,64,1,1) GN,LR	Conv(1,16,5,1) GN,LR	Conv(384,64,1,1) GN,LR
Layer 2	Conv(32,32,3,1) MP(2,2),GN,LR	Conv(128,64,1,1) GN,LR	Conv(64,32,1,1) GN,LR	Conv(16,32,3,2) GN,LR	Conv(64,64,1,1) GN,LR
Layer 3	Conv(32,32,3,1) GN,LR	Conv(64,32,1,1) GN,LR	Conv(32,16,1,1) GN,LR	Conv(32,32,3,2) GN,LR	Conv(64,32,1,1) GN,LR
Layer 4	Conv(32,64,3,1) GN,LR	Conv(32,2,1,1) —	ConvTP(16,16,5,1) GN,LR Conv(16,16,3,1) GN,LR interp(11x11)	Conv(32,32,3,2) GN,LR	Conv(32,6,1,1) —
Layer 5	Conv(64,64,3,1) GN,LR		ConvTP(16,16,5,1) GN,LR Conv(16,8,3,1) GN,LR	Conv(32,32,3,1) GN,LR	
Layer 6	Conv(64,64,3,1) GN,LR		ConvTP(8,4,5,1) GN,LR interp(19x19)	Conv(32,32,5,1) GN,LR	
Layer 7	Conv(64,64,3,1) GN,LR		Conv(4,1,1,1) —	Conv(32,192,3,1) GN,LR	
(x,y,z,c)-in	(42,42,1,3)	(1,1,1,128)	(1,1,1,128)	(25,25,25,1)	(1,1,1,192)
(x,y,z,c)-out	(1,1,1,64)	(1,1,1,2)	(19,19,1,1)	(1,1,1,192)	(1,1,1,6)
# params	139.744	27.138	28.189	393.392	31.238

Table 1. Network Architectures. Building blocks of our architectures are abbreviated as follows: 1.) Conv⟨TP⟩$(c_{in}, c_{out}, kernel, dilation)$ $\widehat{=}$ ⟨Transposed⟩Convolution, 2.) MP$(kernel, stride)$ $\widehat{=}$ MaxPooling, 3.) GN $\widehat{=}$ GroupNorm, 4.) LR $\widehat{=}$ LeakyReLU, 5.) interp$(width, height)$ $\widehat{=}$ upscaling to the specified dimensionality

3 Experiments & Results

To evaluate our contributions, we compare the two self-supervised pre-training schemes on a few-shot CT segmentation task with respect to Dice scores.

Dataset: We perform experiments on the VISCERAL Anatomy3 data [13] using the contrast-enhanced thoracoabdominal scans (training: 63 unlabeled silvercorpus scans, testing: 19 expert labeled scans; leaving out corrupted scans). After resampling to isotropic voxels of size $1.5mm^3$, we crop all images to roughly the same region containing 6 target structures (liver, spleen, left/right kidney, left/right psoas major muscle) - yielding image sizes of 243x176x293 (LR-AP-SI).

Training: In general, we share the same setting for the two compared self-supervision pre-training schemes. Using an Adam optimizer with an initial learning rate of $5 \cdot 10^{-5}$ and a batchsize of 8, we train each method on 800,000 random batches. Each method outputs a feature descriptor of length 192 per position. Details with respect to the network architectures of the different approaches can be found in Table 1. Note that all CNN for descriptor extraction have $\approx 400k$ parameters, have comparably powerful auxiliary task CNNs and are trained as Siamese networks.

In addition to the two self-supervised learning approaches presented in Sec. 2 we consider two additional baselines in our experiments and an ablation study to our proposed HEATMAP ■ method.

Xavier2D: In order to assess the necessity to train the D2D-CNNs in the first place, we also extract 2.5D descriptors with network weights initialized by the Xavier method and without any subsequent training.

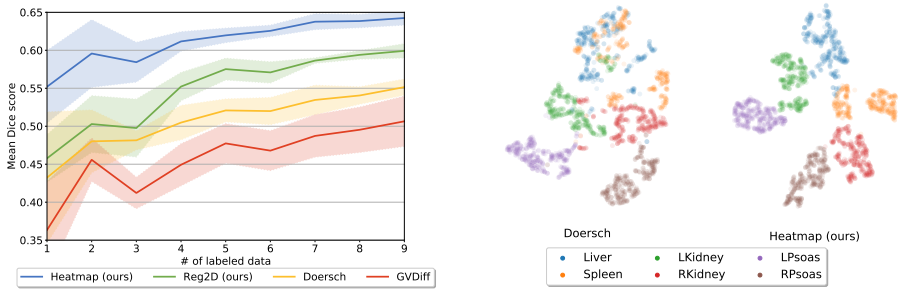


Fig. 2. *Left*: Mean Dice scores for different methods over an increasingly number of labeled testdata. *Right*: t-SNE plots visualizing the more clearly separated feature descriptor clusters for our proposed HEATMAP ■ method compared to DOERSCH ■.

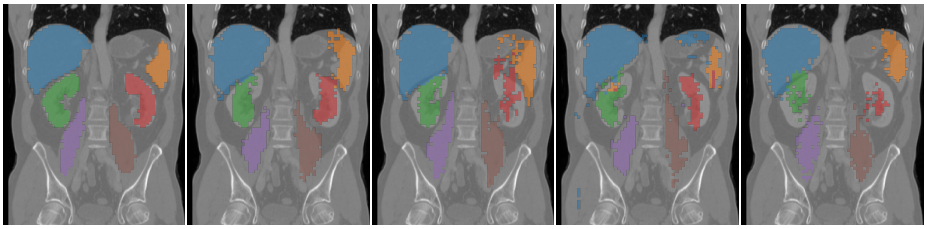
GVDiff ■: As comparison to 'classical' hand-crafted methods, we extract greyvalue difference features (cf. [1]) with a 3D random pattern sampled from a Gaussian distribution with standard deviation 0.4 - i.e. comparable to the receptive fields of the CNN-based methods.

Reg2D ■: To examine the influence of the heatmap approach, we alter our proposed auxiliary learning task to a direct regression of the displacements (Δ_1, Δ_2). In contrast to combining the D2D-CNNs with *Heatmap CNNs*, we use the L1-Loss as penalty to train *2 Param* CNNs that do not reconstruct any spatial information and only operate on the 1D descriptor signals (see Table 1 for architectural details).

3.1 Results

We evaluate all 5 extracted descriptors on 19 datasets, with manual expert annotations. We perform two-fold cross validation (splits: 1-10, 11-19) and examine the influence of an increasing number of labeled datasets (one-shot,2,3,...,9). We predict the organ label at every 4th voxel (192,720 positions per image) - based on an approximate k-Nearest Neighbor (kNN) search using the Vantage Point Forest Method introduced in [5] with $k = 21$ and 15 trees - and compute the resulting Dice as indirect measure of descriptor expressiveness. Note, that we do *not* employ any finetuning strategies to this segmentation task that would require additional GPU-DCNN-training hours - instead building & evaluating the kNN-Forests takes only a few seconds per scan.

Table 2 provides the mean scores for all 6 considered organ structures given a labeled patient database of size 9. Qualitative results with respect to the organ segmentation task are shown in Fig. 3 for a 2D slice of a patient. Fig. 2 (left) shows the mean Dice scores over all organ structures for all patients and folds with an increasing number of available labeled datasets for the kNN classification. Overall, our proposed the HEATMAP-approach performs best and achieves



(a) Expert (b) Heatmap (c) Reg2D (d) Doersch (e) GVDiff

Fig. 3. Segmentation visualization of different approaches for a 2D slice of a patient.

a one-shot segmentation accuracy of $\approx 55\%$ average Dice score. Our alternative approach Reg2D achieves the second highest accuracy and also outperforms Doersch, the straight-forward 3D extension of [2]. With both auxiliary task implementations, our proposed new 2.5D scheme demonstrates its usefulness as self-supervised pre-training scheme for 3D image data. Interestingly, we also outperform [10], which proposed a sophisticated dual CNN architecture specifically designed for one-shot segmentation and achieved 52.6% Dice accuracy on the same dataset.

Visualizing the extracted features using an unsupervised t-SNE embedding from the same foreground positions in Fig. 2 (right) (no labels provided during training) shows the discovery of very clean and separable clusters for individual structures using our Heatmap method compared to Doersch - supporting our hypothesis that leveraging a larger context is of great importance in self-supervised learning.

4 Conclusion

We have presented a novel self-supervised pre-training strategy to effectively leverage inherent 3D information from abundant unlabeled medical volumes. Inspired by the method proposed in [2] for 2D natural images, we designed a new context prediction task that takes explicit advantage of the third image dimension and uses nearly planar subvolumes to train an auxiliary task for continuous and small axial offset prediction between patches. This process, which is repeated

Experiment	Liver	Spleen	LKidney	RKidney	LPsoas	RPsoas	Mean
Heatmap (ours)	85.3	65.7	66.3	53.5	50.4	65.6	**64.2 ± 2.9**
Reg2D (ours)	81.4	54.0	63.4	51.0	49.0	60.9	60.0 ± 2.9
Doersch	76.9	43.0	59.0	51.2	49.1	52.3	55.2 ± 3.1
GVDiff	80.7	58.2	54.5	43.0	29.0	37.1	50.4 ± 5.0
Xavier	70.1	28.3	17.2	3.3	24.5	27.1	28.4 ± 1.0

Table 2. Mean Dice scores in % over all folds with 9 labeled test images.

for all three orientations enables the convolutional network to intrinsically encode anatomical cues into expressive, pre-trained descriptors. When evaluating our scheme with its extracted features within a few-shot kNN-based organ segmentation task and without any supervised refinement, we obtain a large increase of Dice scores from 55.2% to 65.6% compared to the 3D extension of [2]. Despite the fact that we only trained with spatial relations and perform no fine-tuning, we also achieve state-of-the-art results in accuracy for one-shot-segmentation on a public abdominal CT dataset. In future work a more extensive investigation of the influence of network architectures, including convolution filter hyperparameters will be considered. In addition the use of arbitrarily oriented 2D stacks could further enhance the method and many more medical applications, e.g. image registration could benefit from these pre-trained descriptors.

References

1. Calonder, M., Lepetit, V., Strecha, C., Fua, P.: BRIEF: Binary robust independent elementary features. In: ECCV (2010)
2. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV (2015)
3. Doersch, C., Zisserman, A.: Multi-task self-supervised visual learning. In: ICCV (2017)
4. Ferrante, E., Dokania, P.K., Silva, R.M., Paragios, N.: Weakly-supervised learning of metric aggregations for deformable image registration. IEEE Journal of Biomedical and Health Informatics (2018)
5. Heinrich, M.P., Blendowski, M.: Multi-organ segmentation using vantage point forests and binary context features. In: MICCAI (2016)
6. Jamaludin, A., Kadir, T., Zisserman, A.: Self-supervised learning for spinal MRIs. In: DLMIA (2017)
7. Maier-Hein, L., Ross, T., Gröhl, J., Glocker, B., Bodenstedt, S., Stock, C., Heim, E., Götz, M., Wirkert, S., Kenngott, H., et al.: Crowd-algorithm collaboration for large-scale endoscopic image annotation with confidence. In: MICCAI (2016)
8. Payer, C., Štern, D., Bischof, H., Urschler, M.: Regressing heatmaps for multiple landmark localization using cnns. In: MICCAI (2016)
9. Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., Rabinovich, A.: Training deep neural networks on noisy labels with bootstrapping. ICLR workshop (2015)
10. Roy, A.G., Siddiqui, S., Pölsterl, S., Navab, N., Wachinger, C.: 'squeeze & excite'guided few-shot segmentation of volumetric images. arXiv:1902.01314 (2019)
11. Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M.: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Transactions on Medical Imaging **35**(5), 1285–1298 (2016)
12. Tajbakhsh, N., Hu, Y., Cao, J., Yan, X., Xiao, Y., Lu, Y., Liang, J., Terzopoulos, D., Ding, X.: Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data. ISBI (2019)
13. Jimenez-del Toro, O., Müller, H., Krenn, M., Gruenberg, K., Taha, A.A., Winterstein, M., Eggel, I., Foncubierta-Rodríguez, A., Goksel, O., Jakab, A., et al.: Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: Visceral anatomy benchmarks. IEEE Transactions on Medical Imaging **35**(11), 2459–2475 (2016)

14. de Vos, B.D., Berendsen, F.F., Viergever, M.A., Sokooti, H., Staring, M., Išgum, I.: A deep learning framework for unsupervised affine and deformable image registration. Medical Image Analysis **52**, 128–143 (2019)
15. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV (2016)