High-Particle Simulation of Monte-Carlo Dose Distribution with 3D ConvLSTMs

Sonia Martinot[1,2,3(✉)], Norbert Bus[1], Maria Vakalopoulou[2], Charlotte Robert[3], Eric Deutsch[3], and Nikos Paragios[1]

[1] Therapanacea, Paris, France
sonia.martinot@hotmail.fr
[2] CentraleSupélec, 91190 Gif-sur-Yvette, France
[3] Institut Gustave Roussy, 94805 Villejuif, France

Abstract. Monte-Carlo simulation of radiotherapy dose remains an extremely time-consuming task, despite being still the most precise tool for radiation transport calculation. To circumvent this issue, deep learning offers promising avenues. In this paper, we extend ConvLSTM to handle 3D data and introduce a 3D recurrent and fully convolutional neural network architecture. Our model's purpose is to infer a computationally expensive Monte Carlo dose calculation result for VMAT plans with a high number of particles from a sequence of simulations with a low number of particles. We benchmark our framework against other learning methods commonly used for denoising and other medical tasks. Our model outperforms the other methods with regards to several evaluation metrics used to assess the clinical viability of the predictions. Code is available at https://git.io/JcbxD.

Keywords: Deep learning · Radiotherapy · Recurrent neural network · LSTM · Monte-Carlo · Denoising · Convolutional neural network

1 Introduction

In photon radiotherapy, accurate dose modelling is crucial for the treatment planning process in order to target tumours while leaving surrounding healthy tissues unharmed. For that purpose, Monte-Carlo (MC) methods remain unmatched by conventional algorithms such as pencil beam [9] and collapsed cone [1] in terms of precision. MC methods are based on probabilistic simulation of the behaviour of billions of particles in matter. However, due to the full particle transport modeling, MC methods are still extremely computationally expensive which prevents their extensive clinical adoption. Recent research partially addressed this issue by taking advantage of hardware acceleration and using efficient GPU implementations [4,11].

Modern treatment techniques require the calculation of radiation from complex beam configurations. Intensity-modulated radiation therapy (IMRT) allows only a few distinct gantry angles (the direction of the radiation source) while volumetric modulated arc therapy (VMAT) relies on the continuous movement

M. de Bruijne et al. (Eds.): MICCAI 2021, LNCS 12904, pp. 499–508, 2021.
https://doi.org/10.1007/978-3-030-87202-1_48

of the gantry around the center of the tumor. The latter presents less monitor units, more conformity and its delivery time is faster than IMRT [13], improving patient care. On the other hand, the large irradiated area in VMAT requires more simulated particles and therefore more time to reach high quality dose simulations. This is due to the inherent noise present in MC simulations. As the number of simulated particles increase, the noise that corrupts the underlying true dose decreases. Enabling deep neural network architectures to learn the underlying mechanisms of this causal relationship would lead to a flexible dose prediction framework which would be more resilient to various anatomies and allow efficient denoising of MC simulations of VMAT cases.

In this work, we generalize a recurrent neural network structure called a ConvLSTM cell to cope with three dimensional inputs for denoising of dose maps. Our contributions are threefold: *(i)* we present one of the first deep learning based denoising algorithms on VMAT plans, *(ii)* we introduce a 3D, recurrent and fully convolutional deep learning model that infers low uncertainty MC dose distributions from sequences of high uncertainty and low time complexity MC simulations, *(iii)* we release our VMAT dataset (dose distributions) and code. To the best of our knowledge, this is one of the first time convolutional LSTMs are investigated in a fully 3D setting, which could boost a big range of medical imaging applications, such as [2].

2 Related Work

Artificial intelligence and in particular deep learning, offer promising avenues for the clinical integration of MC methods through a colossal gain in terms of simulation time. Deep learning engines for denoising anatomy specific MC simulations have been proposed recently in the literature. In [12] the authors proposed an encoder-decoder architecture to predict high precision simulations from low precision ones in rectal cancer patients treated with IMRT. Neph et al. [10] used combined UNets [14] coupled with additional CT scans as input to solve the same problem in MR-guided beamlet dose for head and neck patients. Vasudevan et al. [18] investigated Generative Adversarial Networks [3] (GANs) to denoise dose simulations in water phantoms reporting promising results.

In the deep learning domain, convolutional architectures which extract relevant features from spatial information differ from recurrent architectures which exploit sequential correlations. ConvLSTM cells [16] were introduced to take advantage of both spatial and sequential information in a two-dimensional setting. This is a generalisation of Long Short-Term Memory (LSTM) [5] and fully convolutional LSTMs [17] architectures. This study aims to demonstrate that this novel recursive framework harnesses its strength from the sequential nature of its input and its ability to derive correlation between the levels of noisiness induced by the different number of particles simulated in the 3D space.

3 Methodology

3.1 Formulation of the Monte-Carlo Progressive Denoising Task

A MC simulation of radiotherapy dose requires inferring the dose deposited by billions of photons in the human body. The method consists in drawing independent random samples from an unknown distribution by means of sequentially sampling empirical measures that describe the dose deposition of individual photons. Let us denote by $M_{N_i} \in \mathbb{R}_{+}^{d_1 \times d_2 \times d_3}$ the 3D dose volume result of a simulation performed with N_i photons. Since several MC dose simulations for the same patient are independent from each other, the following equation holds:

$$M_{N_i} + M_{N_j} = M_{N_i+N_j}$$

Repeating this addition multiple times allows us to achieve simulations with a high number of samples. We can then assess this cumulative process as a temporal one, where the indices N_i correspond to consecutive time steps. In that case, a dose simulation can be represented by a stochastic variable X_{N_i} that we observe over time, as the number of simulated photons grows. Then, considering a sequence $(X_{N_1}, \ldots, X_{N_T})$ with T observations of that variable, our denoising problem amounts to predicting the most likely observation $X_{N_{T+1}}$ based on the given sequence:

$$X_{N_{T+1}} = \underset{X_{N_{T+1}}}{\operatorname{argmax}}\, p(X_{N_{T+1}} | X_{N_1}, ..., X_{N_T})$$

where p denotes an unknown probability. For our denoising task, an observation of X_{N_i} is the radiotherapy dose delivered to a patient at each time step, i.e. M_{N_i}. Hence, there is a need to exploit both spatial and temporal information of the given sequence before inferring the highly sampled dose. This formulation allows us to exploit the temporal and spatial coherence in the process of simulation.

3.2 LSTM and ConvLSTM Cells

LSTMs are a special type of recurrent neural networks (RNNs), able to exploit long-term temporal dependencies. The major asset of LSTM lies in its memory cell which can accumulate information as a cell state C_t. This cell is modified depending on whether the controlling gates are activated. As the input state i_t enters the LSTM, it is processed by an activation function whose final value can activate the forget gate f_t. When the forget gate is on, the past cell status C_{t-1} may be "forgotten". The current cell state is then propagated to the final state h_t in a way that is determined by the output gate o_t. The notation follows [16].

The memory cell allows LSTMs to circumvent the vanishing gradient problem that occurs in the regular RNN model. However, the LSTM only models 1D temporal information and does not make use of potential spatial information. ConvLSTMs overcome the latter limitation of LSTM. They can process 2D data

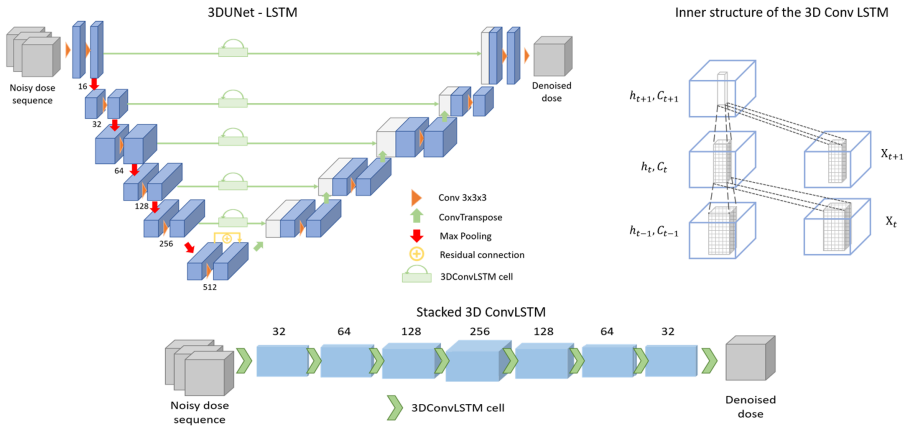

Fig. 1. Proposed architectures and inner structure of the 3D ConvLSTM. The number of output channels after each block appears above or below the layers' output volumes.

such as images by replacing transitional multiplications with convolutions. This innovation allows the model to infer a pixel's next state from its own past status as well as its neighbours.

3.3 Proposed 3DConvLSTM MC Denoiser

3DConvLSTM Cells. As we are considering data that present spatial information in three dimensions, we extended the ConvLSTM framework to deal with temporal sequences of 3D volumes. This can be achieved by using 3D convolutional operators indicated by $*$. In that structure, W_z and b_z in the equations below denote the parameters (filters and bias) of the considered convolutional layers. $\odot$ stands for the Hadamard product and σ for the sigmoid function. The following Eqs. (1–5) describe how gates are activated and states modified:

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \odot C_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + Wcf \odot C_{t-1} + b_f) \tag{2}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \tag{3}$$

$$o_t = \sigma(W_{xo} * X_t + W_{co} \odot C_t + b_o) \tag{4}$$

$$H_t = o_t \odot tanh(C_t) \tag{5}$$

Figure 1 presents the structure of the 3D ConvLSTM. All the states of the 3DConvLSTM cell are initialized with zeros which corresponds to ignorance of the future states. To ensure that dimensions match between the input and the various states inside the cell, padding is applied before convolutions. Therefore, the output of the 3DConvLSTM has the same spatial dimensions as the input.

This extension of ConvLSTM allows processing of medical volumetric sequential data in a fully convolutional manner. Each voxel's future state can be seamlessly predicted using contextual information brought by both temporal and spatial features from its own and its neighbours' past states in all dimensions. In the following subsections we present the two different setups that we used to integrate the 3DConvLSTM cells.

Proposed Model with Stacked 3DConvLSTM Cells: The model consists of 7 3DConvLSTM cells stacked on top of each other, without introducing any spatial downsampling. All convolutional layers in the 3DConvLSTM cells contain $3 \times 3 \times 3$ filters. Figure 1 shows the architecture of the Stacked ConvLSTM model. The spatial dimensions remain equal to that of the input through all propagation.

Proposed UNet with 3DConvLSTM Cells as Skip Connections: We also introduce a model based on the 3DUNet [21] architecture and enhanced with 3D ConvLSTM cells in the skip connections to further extract features at each of the 5 down-sampling steps. Down-sampling is performed using max pooling with kernel size and stride of 2 and comprises two identical convolutional layers. All convolutions have $3 \times 3 \times 3$ filters. The bottleneck has two identical convolutions with a residual connection to further exploit the deep features. The up-sampling blocks have a transpose convolution for up-sampling and a regular convolutional layer for further processing. Each convolutional layer is ended with a LeakyReLU [8] activation function, and batch normalization [6] is used for faster convergence. Details regarding the number of channels at each stage are shown in Fig. 1. This model is trained in the same setting as the proposed model with stacked 3D ConvLSTM cells.

To train all these models, we use a hybrid loss function that adds the Structural Similarity Index Measure (SSIM) [20] and the L1 loss. The parameters of the models are optimized by minimizing the following loss function (6):

$$\mathcal{L} = \sum_{i=0}^{N_{samples}} \left(\left|\left| X_{N_{T+1}}^{(i,estimated)} - X_{N_{T+1}}^{i} \right|\right|_1 + SSIM\left(X_{N_{T+1}}^{(i,estimated)}, X_{N_{T+1}}^{i} \right) \right) \tag{6}$$

where $X_{N_{T+1}}^{(i,estimated)}$ is the model's estimation of the i-th denoised dose volume sample. The SSIM metric is renown for giving a quantitative idea of the perceived quality of an image by measuring the similarity between two images. The L1 loss is well known to help keep track of fine grained details while training a model.

4 Dataset Construction

The patient cohort encompasses 50 patients treated with external beam photon therapy using the VMAT technique. Anatomies are diverse, including 22 pelvic and 28 head and neck cases. The dataset was split to 40, 5 and 5 patients for training, validation and test respectively. Anatomies were distributed as evenly as possible between these sets.

A model for a 6 MeV photon beam with standard fluence of a Varian Truebeam linear accelerator was constructed using schematics and phase space data available in the IEAE phase space database. Each patient plan comprises two VMAT arcs. For each patient, the multileaf collimator (MLC) shapes and gantry angles were extracted from the original clinical plans. MC simulations of such plans were computed with 5×10^8, 10^9, 5×10^9 particles for the noisy dose volumes and 10^{11} particles for the ground-truth dose. The particle transport was simulated using OpenGate [15] with Geant4. The resolution of the simulation was set to 2 mm^3. For the fully sampled dose, the maximum uncertainty in areas within 20%–100% of the dose maximum remained below the clinically accepted 3% threshold. One complete simulation of a 2 arcs VMAT plan required over 4k hours of computation time on CPU without using any variance reduction technique.

4.1 Implementation Details

A patch-based training was implemented by randomly selecting sub-volumes from the 3D input sequences - ground-truth pairs, in areas within 30%–100% of the dose maximum. The patch size was $64\,\mathrm{mm}^3$ subvolumes, i.e. $12.8\,\mathrm{cm}^3$. We used Adam optimizer with learning rate, weight decay, beta1, beta2 and epsilon parameters set to 10^{-5}, 10^{-4}, 0.9, 0.999 and 10^{-8} respectively. The learning rate was reduced by half when the validation loss stagnated, i.e. when difference in loss was inferior to $1e^{-2}$ for more than 200 iterations. The batch size was set to 8. All models were trained for $3 \cdot 10^5$ iterations. The final model we kept was the one that performed best on the validation set.

The input sequence comprises 3 decreasingly noisy dose volumes simulated with 5×10^8, 10^9 and 5×10^9 particles of the same patient case. We use random horizontal and vertical flipping as sole augmentation techniques. The ground-truth was the corresponding highly sampled simulation with 10^{11} particles. Each sample was selected and fed to the model along the axial view.

5 Experimental Results

We compare our method with other commonly used learning based denoising methods in the literature. Our first benchmarking model is a 3DUNet [21] with 5 down-sampling blocks. The second one is Pix2Pix [7], a generative adversarial framework. Pix2Pix has been adapted to a 3D setting. Moreover, since we are

Table 1. Evaluation metrics for the performance of the models on the test set.

Method	SSIM	GPR	L1	# parameters
Inputs 5e9 particles	58.1 ± 0.1	59.1 ± 2.1	0.149 ± 0.050	
3DUNet [14]	80.0 ± 2.4	61.2 ± 2.8	0.088 ± 0.007	10 M
Pix2Pix 3D [7]	55.4 ± 8.6	66.6 ± 14.4	0.102 ± 0.009	120 M
3D BiONet [19]	93.0 ± 0.2	90.6 ± 1.2	0.080 ± 0.001	178 M
Proposed 3DUNet ConvLSTM	64.5 ± 6.1	79.1 ± 1.2	0.037 ± 0.004	36 M
Proposed Stacked 2D ConvLSTM	81.6 ± 3.2	74.1 ± 3.1	0.021 ± 0.003	1.5 M
Proposed Stacked 3D ConvLSTM	**97.9 ± 0.9**	**94.1 ± 1.2**	**0.004 ± 0.001**	5 M

considering smaller data in terms of height and width, we remove one down-sampling block and the corresponding decoding block from the generative model. The adapted generator thus consists of 5 down-sampling blocks, giving a fair comparison with the proposed 3DUNet architecture. Since these models don't handle sequential data, the input is the last volume of the sequence fed to the recurrent architecture, i.e. the least noisy simulation of the sequence. Finally, we also compared with the recently proposed BiONet architecture [19] after adapting it to 3D data and also limiting the number of down-sampling blocks.

Extensive quantitative comparison using the L1 error, SSIM and gamma passing rate (GPR) for each model on the test set are presented in Table 1. We evaluate the GPR criteria with a dose to agreement and tolerance on dose values of 3%/3 mm within 30%–100% of maximum dose. Results show that Stacked 3DConvLSTM outperforms all benchmark models in all metrics while having the lowest number of trainable parameters. We also trained the original ConvLSTMs, on slices of dose volumes. Results in Table 1 reveal that the 2D version still performs better than 3DUNet and Pix2Pix3D with regards to all metrics with only 1.5 million parameters indicating the need of sequential data for this task. Nevertheless, it does not outperform its 3D counterpart nor 3D BiONet. This fact highlights that our 3D model as well as 3D BiONet extract volumetric features that greatly improve the quality of the predictions. Moreover, Stacked 3D ConvLSTM achieves the lowest L1 value and displays GPR scores with standard deviations of 1.2. Although BiONet also shows comparative robustness in its predictions, the quality of the denoised dose volumes remain inferior to that of our proposed models. Another remark stemming from these results is that, despite having a higher SSIM than Pix2Pix3D, 3DUNet's GPR is lower. This might indicate that 3DUNet is able to infer structural coherence in the dose volumes but lacks in precision at a voxel level. In contrast, the proposed recurrent 3DUNet outperforms 3DUNet on the GPR by 18% even though its SSIM score fails to match the 3DUNet.

Figure 2 shows the predictions of the best performing models, namely BiONet and the Stacked 3DConvLSTM, on a test case. Both models reproduce high dose regions well. To further assess the denoising ability of the models, dose profiles are provided in Fig. 3. Both models succeed in smoothing the noise of the

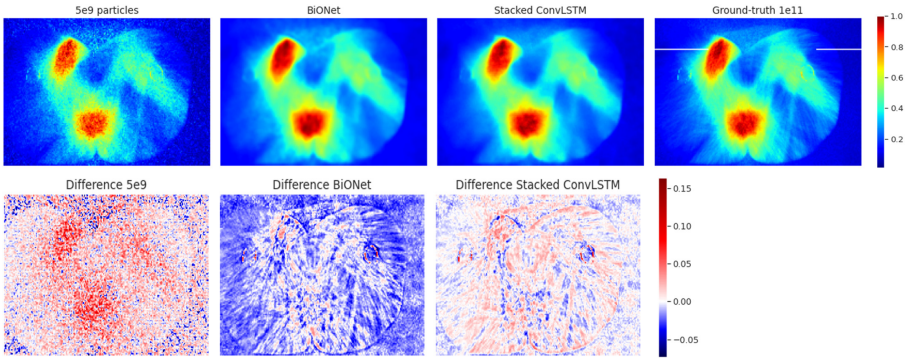

Fig. 2. On the first row from left to left: a single slice of the $5 \cdot 10^9$ dose volume, BiONet's, Stacked 3D ConvLSTM's predictions and ground-truth $1 \cdot 10^{11}$ dose volume. On the second row from right to left error maps for the three different representations.

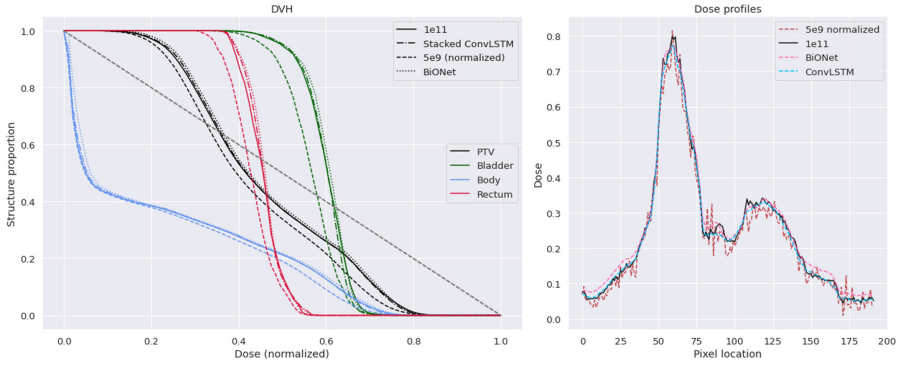

Fig. 3. a) DVH curves, showing that the dose gradients are reproduced faithfully. b) Dose profile along the line indicated on the ground truth image of Fig. 2.

low-simulation input. Error maps between predictions and ground-truth dose associated with BiONet point out that BiONet globally overestimates the dose in low dose gradient regions. Stacked 3DConvLSTM performs better in those regions but underestimates dose in high dose gradient regions where denoising is expected to be more challenging. However, we can notice that both models unfortunately tend to smooth fine details of the Monte-Carlo ground-truth simulation. Figure 3 plots the dose volume histogram (DVH) corresponding to the patient studied in Fig. 2. Both models substantially improve the DVHs towards the ground-truth DVHs. Nevertheless, the DVHs of BiONet indicate that the model still slightly overestimates the dose in voxels, in contrary to the Stacked 3DConvLSTMs.

6 Conclusions

Independently of GPU-accelerated computation, MC simulation time can be further decreased using deep learning based frameworks. The goal of this work is to highlight how considering the MC simulation task as a spatiotemporal problem can be an asset to reach accurate and fast computation of dose. Extensive experiments and comparisons with other state of the art methods highlight the potential of our method. However, the fact that our model does not perform any spatial down-sampling implies that the required GPU memory usage could still be reduced. Achieving high GPR scores while decreasing the computational load could enable real-time Monte-Carlo dose simulation. Future work aims to reduce the number of simulated particles, or in other words increase the level of noise of the input dose volumes.

Acknowledgments. We thank the CEA-TGCC and the TGCC support team for their guidance and the computational resources of the Joliot-Curie supercomputer through a GENCI Grand Challenge. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 880314.

References

1. Ahnesjö, A.: Collapsed cone convolution of radiant energy for photon dose calculation in heterogeneous media. Med. Phys. **16**(4), 577–592 (1989)
2. Gao, Y., Phillips, J.M., Zheng, Y., Min, R., Fletcher, P.T., Gerig, G.: Fully convolutional structured LSTM networks for joint 4D medical image segmentation. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 1104–1108 (2018). https://doi.org/10.1109/ISBI.2018.8363764
3. Goodfellow, I.J., et al.: Generative adversarial networks (2014)
4. Hissoiny, S., Raaijmakers, A., Ozell, B., Després, P., Raaymakers, B.: Fast dose calculation in magnetic fields with GPUMCD. Phys. Med. Biol. **56**(16), 5119 (2011)
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
6. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift (2015)
7. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
8. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models (2013)
9. Mohan, R., Chui, C., Lidofsky, L.: Differential pencil beam dose computation model for photons. Med. Phys. **13**(1), 64–73 (1986)
10. Neph, R., Huang, Y., Yang, Y., Sheng, K.: DeepMCDose: a deep learning method for efficient Monte Carlo Beamlet dose calculation by predictive denoising in MR-guided radiotherapy. In: Nguyen, D., Xing, L., Jiang, S. (eds.) AIRT 2019. LNCS, vol. 11850, pp. 137–145. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32486-5_17
11. Neph, R., Ouyang, C., Neylon, J., Yang, Y., Sheng, K.: Parallel beamlet dose calculation via beamlet contexts in a distributed multi-GPU framework. Med. Phys. **46**(8), 3719–3733 (2019)

12. Peng, Z., et al.: Deep learning for accelerating Monte Carlo radiation transport simulation in intensity-modulated radiation therapy. arXiv preprint arXiv:1910.07735 (2019)
13. Quan, E., et al.: A comprehensive comparison of IMRT and VMAT plan quality for prostate cancer treatment. Int. J. Radiat. Oncol. Biol. Phys. **83**, 1169–78 (2012). https://doi.org/10.1016/j.ijrobp.2011.09.015
14. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
15. Sarrut, D., et al.: A review of the use and potential of the gate Monte Carlo simulation code for radiation therapy and dosimetry applications. Med. Phys. **41**(6Part1), 064301 (2014)
16. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.C.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. arXiv preprint arXiv:1506.04214 (2015)
17. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. arXiv preprint arXiv:1409.3215 (2014)
18. Vasudevan, V., Huang, C., Simiele, E., Yu, L., Xing, L., Schuler, E.: Combining Monte Carlo with deep learning: Predicting high-resolution, low-noise dose distributions using a generative adversarial network for fast and precise Monte Carlo simulations. Int. J. Radiat. Oncol. Biol. Phys. **108**(3), S44–S45 (2020)
19. Xiang, T., Zhang, C., Liu, D., Song, Y., Huang, H., Cai, W.: BiO-Net: learning recurrent bi-directional connections for encoder-decoder architecture. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 74–84. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_8
20. Zhou Wang, Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
21. Özgün Çiçek, Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation (2016)