Conditional 4D Motion Diffusion Models with Masked Observations to Forecast Deformations

Sylvain Thibeault[1], Liset Vazquez Romaguera[1], and Samuel Kadoury[1,2]

[1] MedICAL, Polytechnique Montreal, Montreal, QC, Canada
[2] CHUM Hospital Research Center, Montreal, QC, Canada

Abstract. Image-guided radiotherapy procedures in the abdominal region require accurate real-time motion management for safe dose delivery. Anticipating future 4D motion using live in-plane imaging is crucial for accurate tumor tracking, which enables sparing normal tissue and reducing recurrence probabilities. However current real-time tracking methods often require a specific template and volumetric inputs, which is not feasible for online treatments. Generative models remain hindered by several issues, including complex loss functions and training processes. This paper presents a conditional motion diffusion model treating high-dimensional data, describing complex anatomical deformations. A discrete wavelet transform (DWT) maps inputs into a frequency domain, allowing to select top features for the denoising process. The end-to-end model includes a masking mechanism of deformation observations, where during training, a motion diffusion model is learned to produce deformations from random noise. For future sequences, a denoising process conditioned on input deformations and time-wise prior distributions are applied to generate smooth and continuous deformation outputs from cine 2D images. Lastly, a temporal 3D local tracking module exploiting latent representations is used to refine the local motion vectors around pre-defined tracked regions. The proposed forecasting technique allows to reduce errors by 62% when confronted to a 4D conditional Transformer displacement model, with target errors of 1.29±0.95 mm, and mean geometrical errors of 1.05±0.53 mm on forecasted abdominal MRI.

Keywords: 4D motion · Tracking · Temporal forecasting · Diffusion models · Motion model · Radiotherapy.

1 Introduction

In image-guided radiotherapy (IGRT), the prediction of time-resolved anatomical variations allows to accommodate for latencies in gantry positioning, caused by the accumulation of image reconstruction, tumor tracking and beam modulation steps. During therapy sessions, this implies that when a gating decision is performed, the internal anatomy (i.e. location and overall shape) of a patient has changed. Hence, anticipating organ motion is required to handle system latencies. Current IGRT procedures (i.e. MR-linacs) allow to generate cine images at pre-determined locations, but are constrained to 2D planes and do not offer

real-time 3D information during dose delivery [18]. Forecasting volumes allows to estimate the 3D tumor targets and organ movement during radiation [7].

To capture the motion of organs during radiotherapy, several previous works have exploited recurrent neural networks for temporal prediction [23, 6, 9, 3], notably LSTM's and its different variations, which process sequential data, including video and text [15]. Specifically, for models trained on spatiotemporal changes, Convolutional Long-Short Term Memory [23] were a popular choice for sequential processing. Still, issues such as vanishing gradients hamper their adoption for long-term dependency learning, since the accumulation of errors is common for recursive prediction. Transformers offer improved generalization capabilities compared to recurrent methods [17], but are complex to train.

Diffusion models [8] have gained attention in medical imaging for image generation [5], 3D shape generation and multimodal synthesis for segmentation [16]. These were shown to be more stable to methods based on generative adversarial networks (GAN), particularly for image segmentation and image reconstruction applications, as they are more appropriate to handle important amounts of images. Furthermore, diffusion models can help to generate conditional predictions, as shown for image synthesis processes [19] and for condition-based text generation as show in [12]. In particular, BeLFusion [2] attempted to perform motion prediction using denoising processes performed from features in the latent domain. Still, several steps are required to perform feature disentanglement, but the quality of the pre-trained encoder and decoder is a limitation.

Nonetheless, due to the challenges to integrate diffusion models to render organ shapes from imaging data, determined from the principal modes of variation from surrounding organs and target shapes, this affects the reliability of the generation process. Therefore, their use for motion modeling have not been widespread. Recently, dynamic diffusion models for temporal prediction have mostly been explored for human motion [4] but for specific trajectories, while [10] proposed to enforce rules and physical priors to generate future motion patterns. As in NLP, respiratory motion exhibits a strong reliance on sequential data, which can be used as additional prior to condition future predicted values.

In this work, we introduce a diffusion-based prediction framework that simultaneously learns previous deformations from a series of input volumes and anticipates future representations. The model integrates both the observations and predictions of temporal deformations, learning a motion model that produces organ displacement fields from random noise. At inference time, deformations of the organ are forecasted from an input 2D sequence describing in-plane motions. The model obtains noisy spectrum of deformations by adding noise to the input deformation fields. During the stepwise denoising process, an alternating masking process is applied on the noisy spectrum, allowing to generate different deformation sequence configurations. The process produces predictions which are conditional to the observed sequence and reference volume. A 4D local refinement stage is applied from produced displacement fields, where latent vectors are used to refine motion fields near tumors. The workflow is shown in Fig.1.

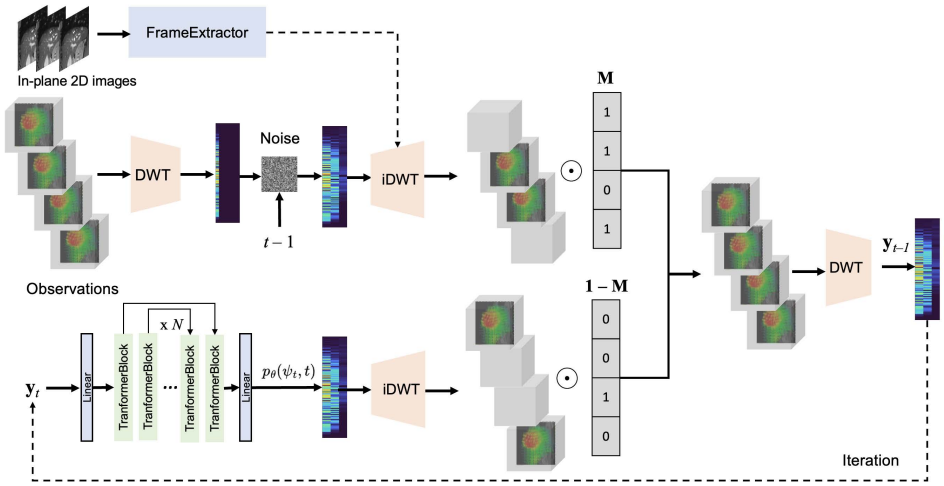

Fig. 1. The proposed forecasting architecture of future liver deformations during image-guided radiotherapy. During training of the diffusion model, a series of deformation vector fields (DVF) are used as input, along with in-plane 2D images, to produce a series of future DVF. Masked deformation observations $\mathbf{M}$ are embedded using the discrete wavelength transform (DWT) and encoded with added noise at time $t-1$, which is conditioned from prior sets of motion distributions $p_\theta(\psi_t, t)$ in feature space Z_f and from the reference volume V_{ref}. The proposed diffusion model produces deformation fields from noisy inputs, which is subsequently used in the inference phase for future predictions $\mathbf{y}_{t-1}$ with masked observation sequence.

2 Methods

2.1 Preliminaries

We denote here a sequence of N deformation observations, denoted as $\phi^{(1:N)} = [\phi^{(1)}, \phi^{(2)}, \ldots, \phi^{(N)}] \in \mathbb{R}^{N \times 3M}$, where $\phi \in \mathbb{R}^{3M}$ is a point in the displacement vector field (DVF) at an observation n, and M is the number of samples in the DVF. For a given series of deformation fields $\phi^{(1:N)}$ with corresponding 2D in-plane images $\mathcal{I} = [I_1, \cdots, I_N]$, the goal of the deformation prediction model is to forecast the subsequent F DVF's, such that $\phi^{(N+1:N+F)} = [\phi^{(N+1)}, \phi^{(N+2)}, \ldots, \phi^{(N+F)}] \in \mathbb{R}^{F \times 3M}$, which when applied to a reference volume V_{ref}, allows for tracking of the internal anatomical motion.

2.2 Frame-wise feature extractor

Using the input 2D in-plane images $\mathcal{I}$, feature vectors are produced using a temporal encoder [17], serving as visual tokens for the diffusion model predictor. The feature encoder receives as input the concatenation of sequential images using a constant baseline image I_{ref}, shared between all the elements of the

series of input images. Using the previous images $I_i|I_{ref}(i = 0, 1, \ldots, N$ observed images$) \in \mathcal{R}^{H_0 \times W_0 \times C_0}$, where H_0, W_0, C_0 and H, W, C denote the baseline and last image height, width, and channels, the module produces a series of feature maps $Z_p \in \mathcal{R}^{m \times H \times W \times C}$ with m observations. Furthermore, the module obtains the ground-truth series $I_{t+j}|I_{ref}(j = 1, 2, \ldots, F)$ of next image representations used as input to the diffusion model, with the generated feature map Z_f exploited as a prior, where at inference time, only the previous frames are used as input.

2.3 Discrete Wavelength Transform

In this study, we use the Discrete Wavelength Transform (DWT) to forecast and produce deformation fields of the liver. The DWT process obtains the temporal characteristic of the deformation sequence, both at resting and periodic phases. The DWT allows to generate smooth deformations across sequences, and is used to train the diffusion model. Provided a $(N + F)$ deformation sequence ϕ, we map the series into the DWT space with the function:

$$\psi = \mathrm{DWT}(\phi) = \mathbf{D}\phi \tag{1}$$

with $\mathbf{D} \in \mathbb{R}^{(N+F)\times(N+F)}$ as the basis of the DWT, and $\psi \in \mathbb{R}^{(N+F)\times 3M}$ representing the coefficients of the transform. Due to the nature of the DWT transformation in orthogonal space, the DWT can capture the deformation sequence through an inverse DWT, defined as iDWT:

$$\phi = \mathrm{iDWT}(\phi) = \mathbf{D}^{\mathrm{T}}\psi. \tag{2}$$

Due to the smooth and continuous nature of organ deformations, we only use the first K rows of the coefficient matrices $\mathbf{D}$, which simplifies the forward and inverse transformations of the DWT by omitting the high-frequency components of the mapping, thus reducing computational cost.

2.4 Conditional diffusion model training

The DWT mapping presented above is applied on the full deformation sequence defined as $\phi \in \mathbb{R}^{(N+F)\times 3M}$ on the top K frequency data in order to generate $\psi_{\mathbf{0}} \in \mathbb{R}^{K\times 3M}$, with $\psi_{\mathbf{0}} = \psi$. The spectrum with added noise obtained at time t is determined by a new parameterization of ψ, so that $\psi_t = \sqrt{\alpha_t}\psi_{\mathbf{0}} + \sqrt{1 - \alpha_t}p$ with α as the variance parameters that are pre-defined before training, where $\alpha_t = \prod_{i=1}^{t} \alpha_i, \alpha_i \in [0, 1]$ and p follows a normal Gaussian distribution.

Then for the noise prediction network, we use a series of N attention blocks (Fig. 1) with linear layers at the beginning and end of the network, producing at each timestep t, a predicted noise defined by $p_\theta(\psi_t, c, t)$, where the parameters are optimized using the following noise prediction loss term:

$$\mathcal{L} = \mathbb{E}_{p,t}\Big[\big\|p - p_\theta(\psi_t, \boldsymbol{c}, t)\big\|^2\Big]. \tag{3}$$

This loss function is the primary objective term for the proposed pipeline to train the diffusion deformation model in an end-to-end fashion, through a minimization of the loss $\mathcal{L}$ that estimates the difference between the expected and observed distributions. The loss is augmented with a conditional parameter $\boldsymbol{c}$ that describes the context of the anatomical image V_{ref}, and is associated with features maps Z_f, obtained from previous observations on the 2D cine images.

2.5 Masked observation inference

To infer a sequence of future deformations, a series of previous observations is provided to the trained model, where T denoising steps are used to generate the series of deformations. We introduce here a method producing a sequence of observations which are padded with future observation data. This is then projected to the DWT space (denoted as ψ), which is added with noise in order to produce a noisy spectrum defined at timestep $t-1$. Here, prediction are defined as $\psi^d_{t-1} = \sqrt{\alpha_{t-1}}\psi + \sqrt{1-\alpha_{t-1}}\mathbf{z}$, with $\mathbf{z}$ following a 0-centered normal distribution of standard deviation σ_t when $t=1$, and $\mathbf{z}=0$ in other cases. Once noise is added, we apply the following conditional denoising procedure:

$$\psi^d_{t-1} = \frac{1}{\sqrt{\alpha_t}}\Big(\psi_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}} p_\theta(\psi_t, \boldsymbol{c}, t)\Big) + \sigma_t \mathbf{z} \tag{4}$$

that produces the prediction ψ_{t-1} from ψ. We use a masking mechanism, where given the diffusion model, the noisy observation and denoised forecasted spectrum follow a similar distribution. Hence, both spectrum are re-projected to the temporal space using the inverse iDWT, combined with a fusion technique:

$$\psi_{t-1} = \mathrm{DWT}[\mathbf{M} \odot \mathrm{iDWT}(\psi^n_{t-1}) + (1-\mathbf{M}) \odot \mathrm{iDWT}(\psi^d_{t-1})] \tag{5}$$

with $\mathbf{M}$ being the masking vector of $(N+F)$ dimension, allowing to mask out randomly observations and predictions, and $\odot$ representing the Hadamard operation which applies the masking vector to the deformation observations.

2.6 Model-based tracker

Finally, to improve tumor tracking given the predicted DVF by the diffusion model, a condition-based projection operation using the diffusion parameters is applied [17]. It computes a weight map within a localized tumor target location (defined on the baseline reference image prior to IGRT), refining the global DVF.

Provided a tumor location $(x_{ref}, y_{ref}, z_{ref})$, a 3D region of interest is determined around the location. We use the region of interest (ROI) to extract the 3D deformation field forecasted from the diffusion-based deformation model (ϕ^{ROI}), from which the weighted map (S_x, S_y, S_z) is generated for the components in all 3 axis, such that (ϕ^{ROI}_x, ϕ^{ROI}_y, ϕ^{ROI}_z).

For any timepoint, a refined 3D DVF associated to a specific plane i is defined, such that ($\hat{\phi}^{ROI}_i$) is obtained by combining all multiplied elements of a predicted motion field (ϕ^{ROI}_i) with parameters (S), i.e., $\hat{\phi}^{ROI} = \|(S_x \times \phi^{ROI}_x, S_y \times \phi^{ROI}_y, S_z \times \phi^{ROI}_z)\|$. Attention maps for a plane i is obtained with [14]:

$$S_i = \sigma_2(\sigma_1(W_c c + W_\phi \phi^{ROI}_i)W_s) \tag{6}$$

with c representing the latent representation of the deformation network, σ_1 and σ_2 the ReLU and sigmoid activations, while W_c, W_ϕ and W_s are linear transformations. The objective when training the tracking module is the minimization of the dissimilarity between predicted ROI and the ground-truth location of interest, in addition to reducing the difference in the prediction and measures from deformation fields around the tumor target area.

3 Results and discussion

A 4D-MRI dataset of 30 radiotherapy patients was used in this study, each providing consent under an IRB-approved protocol. Each patient had 20-minute acquisitions which were obtained free-breathing, with a series of sagittal planes obtained on a 3T Philips Ingenia scanner, with a 2D T2-weighted balanced turbo field echo sequence. The acquisitions of the in-plane frames were centered around the liver lobe and navigator slices were made following an interleaved scheme and subsequently sorted to produce time-resolved 4D datasets [20] with a spatial resolution of 1.7x1.7 mm^2, 3.5 mm slice spacing, and a temporal resolution of 350ms. Each patient had 85 separate sequences with an associated 2D navigator exhibiting various motion frequencies and amplitudes, producing variations in the inter-cycle motion, thus improving model robustness. Providing a series of 2575 volumes per patient across multiple cycles (totaling close to 42000 total volumes), we used a leave-one-out validation scheme, where 29 patients were used for training the remaining case for testing. In this work, DVF between pairs of volumes were pre-computed using a VoxelMorph [1] model trained on liver MRI. This model was used to register liver images between several phases, demonstrating robustness to breathing patterns and deformations [18].

The dataset included between 240 and 400 breathing cycles across the patient acquisitions. We trained the diffusion model on the 4D-MRI dataset using a 1000-step diffusion model and use a 100-step sampling procedure with DDIM [21], which fine-tunes the model on the trained model. Variance scheduling is used here by the Cosine scheduler for the integration of DWT [13]. The feature encoder is composed of a stack of convolutional layers with channels of [64, 128, 256] for the in-plane image forecasting. We used a 8-layer network for the noise prediction module, including linear transformation blocks for the initial mappings, and 4-layer linear transformation modules. Here, $K = 15$, the dropout rate is 0.2 and the latent dimensionality is 512. The parameters of the network were optimized using AdamW with a learning rate of 10^{-3}. For model fine-tuning, the rate was adapted to 10^{-5} and progressively reduced after 3 epochs without improvements in the validation loss. Training was done in PyTorch with a batch size of 10, on a NVIDIA Tesla A100-80GB GPU.

We first evaluated the model's prediction capability based on landmark localization errors throughout different stages of the respiratory cycle. The tumor target area was identified from a trained radiologist. Table 1 presents the target registration errors from the tumor target regions, comparing the proposed model the recent spatiotemporal predictive methods, including MotionDiff [22], LMC [11] and a Transformer-based approach [17], which were trained with similar

Table 1. Target registration errors (TRE) (in mm) from the leave-one-out cross-validation, compared to state-of-the-art methods and ablation experiments. Metrics were extracted at different horizons (h), with 350ms intervals. Results capture the performance across all respiratory phases. Results are mean ± std [P_{90}]. DiffM: proposed diffusion model. C: Conditional prior. MO: Masked observation. T: Tracker module.

Method	TRE (h = 350ms)	TRE (h = 700ms)	TRE (h = 1050ms)
Initial motion	7.28±3.59 [8.87]	7.08±3.32 [8.1]	6.77±3.26 [8.17]
LMC [11]	4.55±2.13 [6.36]	4.49±2.07 [6.31]	4.38±2.02 [6.27]
MotionDiff [22]	3.63±2.28 [5.95]	3.60±2.19 [6.02]	3.55±2.14 [5.92]
Trans4DMoco [17]	2.13±1.76 [3.66]	2.11±1.73 [3.61]	2.08±1.72 [3.74]
DiffM	2.30±1.71 [3.72]	2.28±1.74 [3.62]	2.25±1.82 [3.60]
DiffM+C	1.87±1.32 [3.21]	1.85±1.31 [3.26]	1.82±1.29 [3.29]
DiffM+MO	1.83±1.31 [3.25]	1.84±1.30 [3.31]	1.80±1.28 [3.30]
DiffM+C+MO	1.57±1.23 [2.85]	1.50±1.21 [2.81]	1.49±1.23 [2.82]
DiffM+C+MO+T	**1.31±1.02 [2.17]**	**1.29±0.93 [2.14]**	**1.28±0.92 [2.11]**

conditions to the proposed model. Due to the fact the generated volumes had a temporal resolution of 350ms, the total horizon for prediction was 1050ms. The table also presents results from ablation experiments, where the conditional factor (C), masked observation (MO) and the final tracker (T) components were each assessed. Experiments demonstrate a significant decrease in TRE errors when integrating all proposed components (C, MO and T), with overall errors of 1.29±0.95, when compared to the baseline diffusion model. Furthermore, it is crucial to note that the tracking regions fall outside the in-plane images. In fact in several scenarios, a decrease in the mean error will be observed as the temporal horizon increases. Since metrics are extracted from volumes, predictions tend to be similar where the progressive error increase will not be apparent.

The model yielded geometrical errors, measured between predicted and GT motion fields based on the entire anatomy, of 1.05 ± 0.53mm, compared to 1.54 ± 0.96mm for a Transformer-based approach[17]. Fig. 2 presents the evolution of the overall geometrical accuracy of the prediction models across the different points in the respiratory cycle. Finally, Fig. 3(a) illustrates sample tracking results from 2 test cases, obtained around the tumor target area, in contrast to the forecasted tumor target, while Fig. 3(b) describes motion trajectories with irregular patterns. The proposed approach produces images similar to the ground-truth, where one can notice that the proposed model produces predictions close to the actual MRI acquisitions. Furthermore, a decrease in the errors can be seen when including the tracker method, as opposed to other predictive models based on the Transformer-based approach.

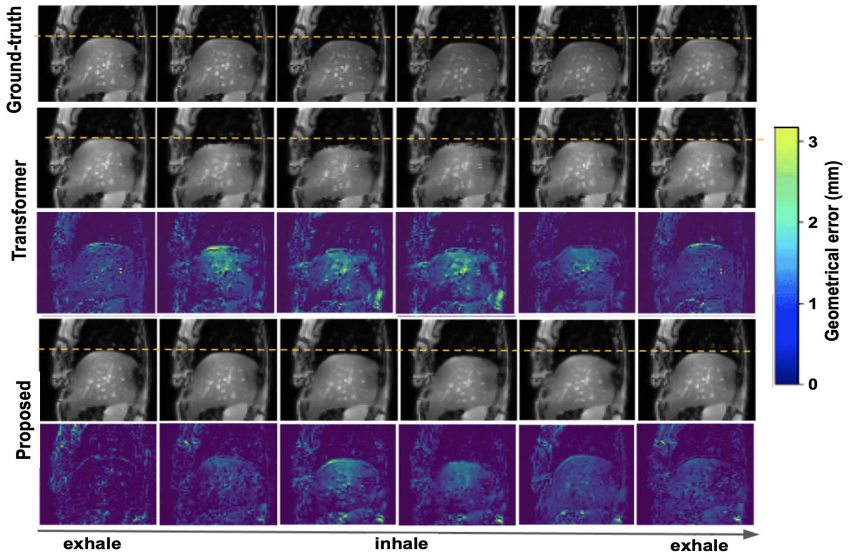

Fig. 2. Series of predicted 4D-MRI sequences and geometrical error maps (in mm) obtained with the proposed and Transformer-based [17] predictors across a respiratory cycle. The dashed lines describe the diaphragm position across the breathing cycle.

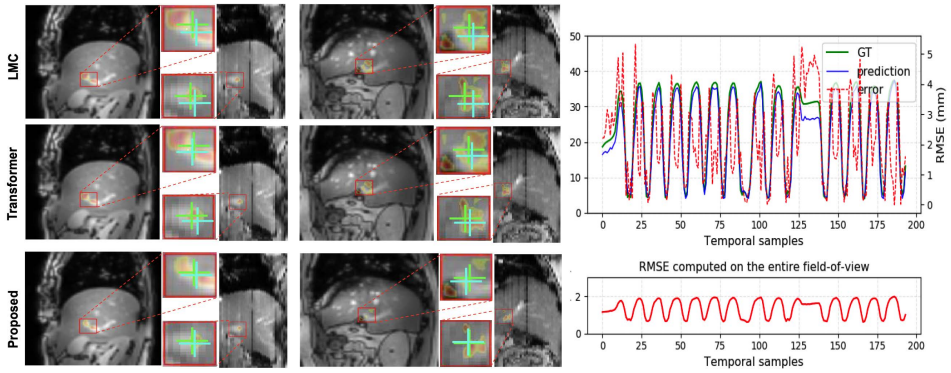

Fig. 3. (a) Forecasted volumes and local tumor locations from the diffusion model with the local tracker, with comparative models on two sample patients (columns). Tumor targets are highlighted in red boxes, green cross-hairs show real tumor locations, while predictions are shown in cyan. (b) Ground-truth and predicted 3D motion trajectories across several breathing cycles.

4 Conclusion

In this paper, we proposed a forecasting model anticipating future organ deformations during IGRT procedures from free-breathing cine MR images of the

liver. The diffusion model captures the main deformation modes from a discrete wavelength transform, while improving model generalizability with a masking mechanism that allows to augment the observational sequences of DVF. Smoothness consistency is integrated within the diffusion model's training process as a conditional prior which allows to produce anatomically consistent sequences. The forecasting model yielded results comparable to ground-truth 4D-MRI observed from separate free-breathing sequences. Conditional diffusion models capture the anatomical variations in organ appearance, thus helping to adapt dose delivery during abdominal RT sessions. This proof of concept shows sufficient accuracy can be achieved in comparison to previous methods, but with higher inference times. Latent Consistency Models (LCMs) is a promising avenue to produce images around 150ms (acceptable for IGRT), as opposed to 10 seconds with vanilla Stable Diffusion. Future studies will consist of assessing the technique's predictive robustness and accuracy in the context of a multi-center study and apply the model for online motion management.

Acknowledgement. Supported by NSERC/Mitacs Alliance Grant 585700-23.

Disclosure of Interests. S Kadoury has received research grants from Varian Medical Systems. Other authors have no competing interests to declare that are relevant to the content of this article.

References

1. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging **38**(8), 1788–1800 (2019)
2. Barquero, G., Escalera, S., Palmero, C.: Belfusion: Latent diffusion for behavior-driven human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2317–2327 (2023)
3. Castrejon, L., Ballas, N., Courville, A.: Improved conditional vrnns for video prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7608–7617 (2019)
4. Chen, L.H., Zhang, J., Li, Y., Pang, Y., Xia, X., Liu, T.: Humanmac: Masked motion completion for human motion prediction. arXiv preprint arXiv:2302.03665 (2023)
5. Dar, S.U.H., Ghanaat, A., Kahmann, J., Ayx, I., Papavassiliu, T., Schoenberg, S.O., Engelhardt, S.: Investigating data memorization in 3d latent diffusion models for medical image synthesis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 56–65. Springer (2023)
6. Denton, E., Fergus, R.: Stochastic video generation with a learned prior. In: International Conference on Machine Learning. pp. 1174–1183. PMLR (2018)
7. Henke, L., Kashani, R., Robinson, C., Curcuru, A., DeWees, T., Bradley, J., Green, O., Michalski, J., Mutic, S., Parikh, P.: Phase i trial of stereotactic mr-guided online adaptive radiation therapy (smart) for the treatment of oligometastatic or unresectable primary malignancies of the abdomen. Radiotherapy and Oncology **126**(3), 519–526 (2018)

8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
9. Hsieh, J.T., Liu, B., Huang, D.A., Fei-Fei, L., Niebles, J.C.: Learning to decompose and disentangle representations for video prediction. arXiv preprint arXiv:1806.04166 (2018)
10. Jiang, C., Cornman, A., Park, C., Sapp, B., Zhou, Y., Anguelov, D., et al.: Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9644–9653 (2023)
11. Lee, S., Kim, H.G., Choi, D.H., Kim, H.I., Ro, Y.M.: Video prediction recalling long-term motion context via memory alignment learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3054–3063 (2021)
12. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
13. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021)
14. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
15. Oprea, S., Martinez-Gonzalez, P., Garcia-Garcia, A., Castro-Vargas, J.A., Orts-Escolano, S., Garcia-Rodriguez, J., Argyros, A.: A review on deep learning techniques for video prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
16. Özbey, M., Dalmaz, O., Dar, S.U., Bedel, H.A., Özturk, Ş., Güngör, A., Çukur, T.: Unsupervised medical image translation with adversarial diffusion models. IEEE Transactions on Medical Imaging (2023)
17. Romaguera, L.V., Alley, S., Carrier, J.F., Kadoury, S.: Conditional-based transformer network with learnable queries for 4d deformation forecasting and tracking. IEEE Transactions on Medical Imaging (2023)
18. Romaguera, L.V., Mezheritsky, T., Mansour, R., Carrier, J.F., Kadoury, S.: Probabilistic 4d predictive model from in-room surrogates using conditional generative networks for image-guided radiotherapy. Medical Image Analysis p. 102250 (2021)
19. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
20. von Siebenthal, M., Szekely, G., Gamper, U., Boesiger, P., Lomax, A., Cattin, P.: 4d mr imaging of respiratory organ motion and its variability. Physics in Medicine & Biology **52**(6), 1547 (2007)
21. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
22. Wei, D., Sun, H., Li, B., Lu, J., Li, W., Sun, X., Hu, S.: Human joint kinematics diffusion-refinement for stochastic motion prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 6110–6118 (2023)
23. Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Advances in neural information processing systems. pp. 802–810 (2015)