Generative Learning Approach for Radiation Dose Reduction in X-Ray Guided Cardiac Interventions

Fariba Azizmohammadi[1], Iñaki Navarro Castellanos [2], Joaquim Miró[2], Paul Segars[3], Ehsan Samei[3],Luc Duong[1]

[1]Interventional Imaging Lab, Department of software and IT engineering, École de technologie supérieure, 1100 Notre-Dame West, Montreal H3C 1K3, Canada

[2]Department of Pediatrics, CHU Sainte-Justine, Montreal H3T 1C5, Canada

[3]Department of Radiology, Carl E. Ravin Advanced Imaging Laboratories, Duke University Medical Center, Durham, NC, USA

Abstract

Background: Navigation guidance in cardiac interventions is provided by X-ray angiography. Cumulative radiation exposure is a serious concern for pediatric cardiac interventions.

Purpose: A generative learning-based approach is proposed to predict X-ray angiography frames to reduce the radiation exposure for pediatric cardiac interventions while preserving the image quality.

Methods: Frame predictions are based on a model-free motion estimation approach using a Long Short Term Memory (LSTM) architecture and a content predictor using a Convolutional Neural Network (CNN) structure. The presented model thus estimates contrast-enhanced vascular structures such as the coronary arteries and their motion in X-ray sequences in an end-to-end system. This work was validated with 56 simulated and 52 patients' X-ray angiography sequences.

Results: Using the predicted images can reduce the number of pulses by up to 3 new frames without affecting the image quality. The average required acquisition can drop by 30% per second for a 15 frame per second acquisition. The average Structural Similarity Index Measurement (SSIM) was 97% for the simulated dataset and 82% for the patients' dataset.

Conclusions: Frame prediction using a learning-based method is promising for minimizing radiation dose exposure. The required pulse rate is reduced while preserving the frame rate and the image quality. With proper integration in X-ray angiography systems, this method can pave the way for improved dose management.

This is the peer reviewed version of the following article: Azizmohammadi, F, Navarro Castellanos, I, Miró, J, Segars, P, Samei, E, Duong, L. Generative learning approach for radiation dose reduction in X-ray guided cardiac interventions. Med Phys. 2022; 1- 11, which has been published in final form at https://doi.org/10.1002/mp.15654.

I. Introduction

Congenital Heart Disease (CHD) affects 1% of the population and is the most common type of birth malformation worldwide[1]. Patients with CHD are exposed to substantial amounts of ionizing radiation from diagnostic and treatment procedures[2]. In recent years, the number of complex, long-duration pediatric cardiac interventions has risen significantly. Consequently, the risks associated with radiation exposure among patients have also increased, which is why solutions must be found to reduce the radiation dose to as low as reasonably achievable (ALARA) while maintaining the required image quality[3]. Minimizing radiation exposure in pediatric cardiology is paramount in interventional cardiology. Patients are subjected to either deterministic outcomes, such as skin necrosis, which is most commonly related to tissue rebounds, or stochastic effects, such as an increased risk of radiation-induced cancer and brain tumors[4]. Moreover, complex CHDs must be catheterized repeatedly, thereby increasing the risk of radiation-induced cancer not only for patients but also for medical staff[5]. Radiation exposure is, therefore, a major concern for pediatric populations, and determining the optimal dose for each patient is a highly relevant research topic in pediatric cardiology.

I.A. Radiation dose reduction in X-ray angiography

Currently, X-ray angiography is widely accepted for minimally invasive interventions and provides adequate spatial and temporal image resolution. Fluoroscopy and fluorography are the two main fluoroscopically-guided intervention modes in X-ray imaging. In fluoroscopy mode, the X-ray images are generated instantaneously and continuously to observe moving objects by capturing the motion. The images in this mode are not recorded and used to navigate the medical devices to specific locations within the patient in real-time. Fluorography mode requires a higher radiation exposure to generate and record high-resolution images for interpretation after the termination of the exposure[6]. The required radiation dose for each acquisition mode is a function of the required image quality, the patient's size, and the time required to perform the procedure. Fluoroscopy time comprises the total time spent using fluoroscopy for image acquisition and is considered as one of the effective parameters for the final patient dosage[7].

Previously, conventional analog X-ray equipment was used to deliver X-ray energy in a continuous dose. Recently, some strategies are applied to mitigate the radiation dose to the patients such as using the lowest possible fluoroscopic dose rate during live fluoroscopy, use of low frame rates (if possible), and use of multiple short fluoroscopic exposures instead of keeping the fluoroscope on continuously and minimizing the beam-on time for the fluoroscopy imaging[8 6].

Modern X-ray systems are equipped to deliver energy in pulses that can be adjusted to 7.5, 10, 15, and 30 frames per second (fps). In pulsed fluoroscopic imaging, the X-ray beam is switched on and off for every fluoroscopic frame, and thus, the pulse width, or time duration of each frame, is lower than the time required in continuous fluoroscopy

imaging. This allows reducing the fluoroscopy time by replacing the continuous exposure with a pulsed beam delivery. However, images are temporally averaged and moving objects look unsharp and flicking. A sequence of pulsed images, including moving objects, appears more continuous and less flickering at high pulse rates or frequencies based on the critical flicker frequency. At low frame rates, gap-filling by replicating each acquired frame multiple times is applied to avoid flicker and minimize blurriness of moving targets. The term frame rate describes the number of frames that are generated per second while the term pulse rate, refers to the output of the fluoroscope, specifically the number of bursts of radiation that are emitted per second[9].

Reducing the pulse rate during complex invasive cardiovascular procedures results in a considerable reduction of the total energy and the patient dose required for X-ray imaging[10]. The average required dose rate scales as the square root of the frame rate, with an equal noise perception for the operator's eyes in pulsed fluoroscopy imaging[9 11]. Hence, if the frame rate is reduced from 15 fps to 7.5 fps, the required dose rate is reduced by 30%, while doubling the frame rate from 15 fps to 30 fps increases the required dose rate by about 40%[9 11]. One common approach for reducing the fluoroscopy time in X-ray fluoroscopy systems, involves the last image hold technique[12].

I.B. Relationship between motion estimation and the dose reduction for cardiac interventions

To keep the radiation dose as low as possible during the diagnostic and interventional procedures, motion compensation and prediction techniques are required to reduce potential misinterpretations caused by motion while preserving the image quality. Cardio-respiratory motion prediction has always been preferred in cardiac applications as it facilitates more accurate navigation procedures.

Deep learning architectures such as Recurrent Neural Network (RNN) models are popular in cardiac imaging and in predicting the cardio-respiratory motion in diagnostic and interventional imaging processes[13 14 15]. In these approaches, motion features (temporal and spatial) are extracted from image frames and memorized by the RNN model to predict upcoming images. However, predicting and generating realistic images and motion in an end-to-end system continues to present issues using existing models. Generative adversarial networks (GANs) are the tools used for learning deep representations. They can be used for both supervised and semisupervised learning by implicitly modeling high-dimensional data distribution. The main structure of GANs is based on training a pair of networks competing against each other. These two networks are generators and discriminators. The generator is like an art forger and produces realistic synthetic samples like images using a distribution. The discriminator acts as an art expert to distinguish the real sample from the synthetic generated one. These two networks are trained at the same time, allowing them to improve in their respective abilities until the discriminator is unable to tell the real and synthetic samples apart[16]. Recently, GANs have been used as an advent method for video frame prediction. Prediction quality has been improved considerably using GANs, and the

combination with RNNs has made it possible to predict multiple frames as well[17].

I.C. Proposed contribution

The contribution of this study is to predict dynamic X-ray angiography sequences using a generative model. A video frame prediction model is introduced to predict new X-ray angiography frames. We introduced a new loss function to predict the temporal and spatial information of the arteries in angiography sequences. To minimize the vesselness structure differences between the predicted and ground truth images, a multi-scale Hessian-based loss term is added to the loss function presented by Mathieu et al[18]. Then, a predictive RNN-based motion model is trained to estimate the motion and content of single and/or multiple future frame(s) based on previously acquired frames in an end-to-end system.

This work is organized as follows: section $II.A$ describes the data used, section $II.B$ presents the X-ray frame prediction, while $II.C$ presents the model architecture. The results and discussion are presented in section III and section IV, respectively.

II. Materials and Methods

II.A. Data description

We developed and validated our method using both simulated and patient X-ray angiography datasets from Sainte-Justine Hospital. Simulated X-ray sequences generated from realistic XCAT computational phantoms with cardio-respiratory motion[19] were first investigated. The simulated motion included the beating heart and respiratory motions. The simulated dataset includes 56 different patients (32 male and 24 female) and 112 sequences (2 sequences per patient, showing either the left coronary artery or the right coronary artery). All the generated sequences had a length of 75 frames and were acquired at 15 fps. The patient X-ray angiography database comprises 52 different patients with contrasted coronary arteries. This study was reviewed and approved by the Institutional Review Board of Sainte-Justine Hospital. Each patient presents a different number of sequences, with varying lengths. There is a total of 340 sequences, respectively with a minimum and maximum length of 15 and 70 frames. All the data were acquired at 15 fps.

II.B. X-ray angiography frame predictions

In this section, the effects of frame predictions on dose reduction are assessed in terms of the required dose rate and the total fluoroscopy time. The quantitative results of this assessment illustrate that reducing the total fluoroscopy time can have a considerable impact on cumulative radiation exposure reduction.

II.B.1. Assessment of the impact of pulse rate reduction on the total radiation dose reduction

In our approach, we assumed that for any specific frame rate (7, 15, 30, 60 fps) the number of pulses required can be reduced during an X-ray imaging process such that the predicted frames can replace the real X-ray frames. Depending on the X-ray manufacturers, the dose for a given exposure duration is directly related to the pulse rate[20 21] or it can scale as the square root of the frame rate for uniform noise perceived by the operator's eyes[9 11]. In this work, we considered the square root model.

According to this approach, for the same frame rate, a smaller pulse rate (i.e. dose rate) is required since T frames are predicted (Figure 1 (a)). Considering K as the number of previously generated and visited frames and T as the number of predicted frames at each prediction mode, for every $K+T$ frames, T frames are predicted. Thus, the number of pulses required at every second can be reduced by $FR \times (\frac{T}{K+T})$. Hence, the Required Dose Rate (RDR) scales proportionally as:

$$RDR \propto \sqrt{FR \times \frac{K}{K+T}} \tag{1}$$

where the FR is the selected frame rate for the intervention or acquisition (7, 15, 30, 60 fps).

Given the parameter K, which is the number of previously generated and visited frames contributing to the prediction of the new frame/s, the X-ray exposure can pause at each predicting mode and resume in acquisition mode. Assuming t_T as the required time for T frames prediction, t_w as the required time window for $K+T$ acquisitions, and FT as the entire required fluoroscopy time (in seconds), the $\hat{FT}$ is the reduced fluoroscopy time:

$$\hat{FT} = FT - (\lfloor \frac{FT}{t_w} \rfloor \times t_T) + t_r \tag{2}$$

In any time window (t_w), the exposure time is reduced by the amount of time that is required to acquire T frames (t_T). The t_r is the remaining time in the X-ray angiography sequence ($t_r = t_{total} \bmod t_w$, $t_r \in W$) (Figure 1)(b).

Figure 1 (b) is an example showing the difference between conventional continuous fluoroscopy, pulsed fluoroscopy, and our method, in terms of fluoroscopy time. For the pulsed fluoroscopy with frame prediction the $\hat{FT} = \Sigma(t_w - t_T) + t_r = \Sigma f t_i$ while $t_r \in W$. In pulsed fluoroscopy, less energy is exposed as compared to continuous fluoroscopy. In the our approach, the X-ray device is supposed to pause at each prediction mode and resume in each acquisition mode. Thus, the total amount of fluoroscopy required in an X-ray imaging process is reduced.

II.B.2. Cardio-respiratory motion and content estimation in X-ray sequences

The prediction of upcoming frames of a video sequence requires two components, namely, the visual content and pixel displacement through time or motion. Thus, the proposed

network learns the internal representation of image evolution through the sequence based on its content and motion. The model in this work consists of two different encoders, one for the visual content, and a second one for the motion of the image sequence. These two key components need to be decomposed among the images and predicted separately. The motion features are extracted by an RNN-based encoder with Long Short Term Memory (LSTM) and CNN, while the visible content features are only extracted from the last visited image with a CNN-based model. Deep learning methods have been applied successfully for video frame prediction in the literature[22 23 24].

II.C. Model architecture

A generative model is built on an Encoder-Decoder framework. To extract the motion and content features of the images in sequences, a CNN model is used, in combination with an LSTM network. The LSTM cells are used to memorize the periodic aspect of the complex cardio-respiratory motion in the angiography sequences. According to our previous work[13], the LSTM structure is robust enough to deal with different motion patterns in the cardio-respiratory motion signals during prediction. Therefore, an LSTM-CNN combination is used for a general motion estimator. The motion and content are predicted independently, using two encoders. Thus, the spatial and temporal dynamic features of the X-ray images are extracted and encoded separately. The model architecture also includes a concatenating section that combines the outputs of these encoders, as well as a multi-scale residual that is used to avoid information loss before pooling in the network. The last part is the decoder, which reconstructs the predicted images. Figure 2 shows the complete structure of the model.

II.C.1. Motion encoder

A Convolutional LSTM (ConvLSTM) extracts the dynamic features in X-ray sequences. While the pixel-level features are extracted by a Convolutional Neural Network (CNN), the sequential information is provided by the LSTM cells in the motion encoder. The motion encoder captures the local motions from one frame to the next in X-ray sequences. The cardio-respiratory movements of the objects (arteries, devices, catheters, wires, stents, etc.) are predicted directly (without using a surrogate object) and independently in the sequences.

The original presented motion encoder in[22] takes the element-wise image subtraction between (x_t and x_{t+1}) as an input. Since there are background movements in angiography images, the subtraction of original frames includes a lot of artifacts. In our approach, we filtered the input images by vesselness filter first and then subtracted the filtered input images to overcome the artifact caused by the background movement. Thus, the motion encoder tracks only the contrasted arteries' movement to encode the temporal dynamics of transformed images through the sequence (d_t). The output of the motion encoder is a function of filtered time frames subtraction ($x_{v(t+1)} - x_{v(t)}$), memory cell c_t, and d_t.

II.C.2. Content encoder

The content encoder extracts the essential spatial features from the visible contents, such as contrasted moving objects (arteries) and the background (ribs, bones, and devices) in the images. It takes the last observed frame x_t as input and encodes the spatial information in the image (CE_t) using a CNN network. The last observed frame has the most recent and important information that is required for the prediction of the future frame(s).

II.C.3. Final prediction using the content and motion encoders' outputs

A multi-scale encoder residual is used to compute the residual Res_t at each scale or layer just before the pooling layers of both motion and content encoders. The outputs of both encoders are concatenated and combined with the residual outputs (d_t, CE_t, Res_t) to perform pixel-level predictions in the decoder. These predictions can represent one or more frames in the future. The output of the model[22] is as follows:

$$ME = [d_t, c_t] = f^{motion}(x_{v(t)} - x_{v(t-1)}, d_{t-1}, c_{t-1}) \tag{3}$$

$$CE = f^{content}(x_t) \tag{4}$$

$$Res_t^h = f^{residual}([CE^h, d_{t-1}^h]) \tag{5}$$

$$Output_t = f^{combination}([d_t, CE]) \tag{6}$$

$$\hat{x}_{t+1} = f^{decoder}(Output_t, Res_t) \tag{7}$$

where ME and CE are the motion and content encoder outputs, respectively. Res^h is the residual link at layer h being used to avoid information loss after pooling for each layer, and $Output_t$ represents the combination layer that concatenates the outputs of both motion and content encoders. The new frame is generated as the output of the decoder going through a $tanh(.)$ activation function.

II.C.4. Loss function

A combination of terms (image space and generator loss terms) is minimized in this approach. We adjusted this loss function to predict the cardiac angiography sequences, given that the targets to track and predict in are contrasted arteries. The total loss function is calculated as below considering the α and β as constant weights:

$$L_{Total} = \alpha L_{IM} + \beta L_{GAN} \tag{8}$$

where L_{IM} represents the image space loss as a combination of terms that match the average pixel intensities with L_P, gradient difference to sharpen the predictions and the new added sub-loss called vesselness[25] difference L_{Vss}.

$$L_{IM} = \alpha L_{gdl} + \beta L_P + \gamma L_{Vss} \tag{9}$$

We penalized the difference between the second derivative of the Gaussian filter applied on the predicted and ground truth images with 6 different scales (vesselness σ range: 0.5 - 3 with step size: 0.5). The output of the vesselness filter on the images is the vesselness response image. The second derivatives encode the shape information and the eigenvector corresponding to the smallest eigenvalue is the direction of the blood vessel locally. Hence, the L_{Vss} is applied to minimize the local differences of the predicted and ground truth images, which refer to the shape of the arteries.

The gradient difference term L_{gdl}[18 22] is applied to sharpen the generated images. This term directly assesses the gradient discrepancy of the ground truth and the predictions. The gradient difference between the ground truth image Y and the prediction $\hat{Y}$ is given by:

$$L_{gdl}(\hat{Y}, Y) = \Sigma_{i,j}(||Y_{i,j} - Y_{i-1,j}| - |\hat{Y}_{i,j} - \hat{Y}_{i,j-1}||^{\lambda} + ||Y_{i,j-1} - Y_{i,j}| - |\hat{Y}_{i,j} - \hat{Y}_{i,j}||^{\lambda}) \tag{10}$$

where λ is an integer greater or equal to 1 (here the $\lambda = 1$) and $|.|$ is the absolute value function[18]. The new vesselness difference term L_{Vss} matches the vesselness responses of the predicted and ground truth images.The vesselness difference between the ground truth image Y and the prediction $\hat{Y}$ is given by:

$$L_{Vss}(\hat{Y}, Y) = \Sigma_{i,j}|I_Y - I_{\hat{Y}}| \tag{11}$$

To generate images correctly and avoid having the images being blurred by time, the generator loss in adversarial training L_{GAN} is added to solve the blurriness problem, and induces realism in the image sequences, in addition to sharpening the images[18].

$$L_{GAN} = -\log D([x_{1:t}, G(x_{1:t})]) \tag{12}$$

while $D(.)$ represents the discriminator in adversarial training, and $x_{1:t}$ is the input images concatenation. The adversarial discriminator loss (L_d) is defined by:

$$L_d = -\log D([x_{1:t}, x_{t+1:t+T})]) - \log(1 - D[x_{1:t}, G(x_{1:t})]) \tag{13}$$

the concatenation of future ground truth images and all of the predictions are represented as $x_{t+1:t+T}$ and $G(x_{1:t}) = \hat{x}_{t+1:t+T}$ respectively[18 22].

III. Results and validations

The parameters for the X-ray angiography sequences were optimized for both the simulated and patient datasets. The number of iterations was evaluated between 1000 to 1500 for

the simulated dataset and between 2000 and 2500 for the patient datasets. We divided the dataset into two parts: 80% of the dataset for training and 20% for testing. The model was evaluated on each dataset separately. Each sequence was divided into time slots or time windows of minimum $(K + T)$ frames. A single frame was repeatedly predicted at a time, and the prediction was included through the time slot while the previous predicted frame(s) contributed to new predictions. The number of previously generated and visited frames ($K = 7$ to 10) contributing to predict the future frame(s) for the motion encoder was set based on capturing a complete heart cycle in time ($0.8s$ to $1s$) and on the length of the shortest sequences in our dataset. All the parameters and hyperparameters were selected based on different experiments. The hyper-parameters α, β and γ were set to 1, 0.02, 0.01, respectively, based on the experiments.

The quality of the predicted images was reduced by increasing the number of predictions. The visual quality of the predicted images using our method (the vesselness-based MCnet) was assessed as compared to the original MCnet in terms of certain similarity measurement metrics such as Peak Signal-to-Noise-Ratio (PSNR) and Structural Similarity Index Measurement (SSIM) (Table 1 and Table 2). In our experiments, we predicted up to 3 frames with over 60 percent SSIM for both the original and vesselness-based MCnets (Table 1 and Table 2).

According to the experiments, the quality of the predicted images is reduced by increasing the number of predictions. With the simulated data, the first three frames were well-predicted with 24 to 29 PSNR and between 87% to 97% SSIM (Table 1). For the patient dataset the best results refer to K = 10 in which the first 3 predicted frames reach to between 68% and 82% SSIM (Table 2). Our experiments show that, the parameter K must be equal to or greater than the number of frames required to cover a cardio-respiratory cycle. Moreover, the values for the parameter K in our experiments depend on the length of the shortest sequences in our patient dataset such that the $K+T$ must be equal or less than the length of the shortest sequence in our dataset (13 frames). Based on the overall experiments with patient and simulated datasets (Table 1 and 2), the first 3 predicted frames have over 60% SSIM and the vesselness structure is clearly visible. Thus, at each second during the X-ray imaging process, the patients can be exposed to 3 fewer pulses while keeping the same frame rate (15 fps). The required frame acquisition (i.e. pulses) for a 15 fps sequence can drop by 23% to 30% (for K = 10 and, K = 7 respectively), and according to (1), the average required dose rate for 15 fps imaging on every second can be reduced by 0.63 to 0.47, as compared to real acquisition. Figure 4 (a) and (b) show the samples of prediction with K = 7 and 10 respectively and Figure 5 shows the overlay of the manually segmented ground truth arteries (in green) and the predictions.

To evaluate the motion prediction, we applied optical flow to estimate the motion between consecutive predicted frames as well as the ground truth frames. Optical flow is one common approach to detect motion of moving objects in an image sequence and it is defined as the distribution of visible velocities of moving objects in an image. Figure 3 shows the estimated movements between the 4 consecutive frames with optical flow. In the first row, the motion arrows are extracted from the ground truth sequence, and in the second and third rows, the motion arrows are extracted from the predicted images using the vesselness MCnet

and the original MCnet, respectively. The optical flow fields between each moving frame and the previous (source) frame are overlaid by moving frames (F = 7,8,9). The motion vectors in the frames predicted using the vesselness MCnet have mostly the same directions and same intensities in the region of interest (arteries) as the ground truth in all frames, while the intensities and directions of the detected motion vectors are different in the predicted frames using original MCnet.

From the test dataset, we randomly selected 30% of the sequences to evaluate the predicted content of the generated images with K = 7 (visited frames) and T = 3 (predicted frames). Coronary arteries were segmented in three consecutive frames of each selected sequence in both groups (ground truth and predictions) by a trained operator. From the resultant masks, we computed the DICE coefficients and Euclidean distances between the ground truth and the predicted images. Euclidean distance was calculated between the extracted centrelines of the segmented masks. Additionally, we reported results of a conventional gap-filling method (baseline) in the selected dataset (Table 3). The gap-filling method copied multiple times the last visited frame instead of being predicted. The Euclidean distance and DICE coefficients of the predicted images in our method and the ground truth were computed and compared with the Euclidean distance and DICE coefficients of the ground truth and the copied frames used as the gap-filling.

The average computed DICE coefficients (between the ground truth and predicted images) over 3 predictions using our method was 0.78 ± 0.07 while this value for the conventional gap-filling was 0.63 ± 0.05. Table 3 shows the comparison of our approach and gap-filling in terms of the computed Euclidean distances between the centerlines of the ground truth and the predictions. Based on these evaluations, for 3 consecutive frames, the results of the frame prediction with our approach outperform the baseline method (gap-filling).

IV. Discussion

This work presents a novel radiation dose management approach for pediatric interventional cardiology using a generative learning-based video frame prediction approach. This study can also facilitate the navigation of X-ray-guided interventions given the intrinsic motion compensation strategy it has in the frame predictions.

In our approach, a predictive model was introduced rather than an interpolation approach since interpolation methods require both future and former information. In frame prediction using this model, the idea is to extract the cyclic cardio-respiratory motion features from the previous frames and to combine them with the visual content of the last visited frame.

The correlations between spatial and temporal features extracted from the previous frames allow self-supervision of the prediction of single or multiple frame(s) in an end-to-end system. This model can be transferable to adult patients by performing training on clinical data from adults. Additionally, the presented model can be fully adaptive to different patients with distinct respiratory and cardiac motion patterns. Compared to other

video frame applications, X-ray sequences have less inherent uncertainty and variety when it comes to estimating upcoming frames since their grayscale images include limited objects for tracking, and the cardio-respiratory motion is periodic. However, the main challenge with X-ray sequence prediction in comparison to natural video prediction lies in the moving background, which makes motion prediction more complex in the former. In this work, we applied a new loss function and changed the input of the motion encoder using a vesselness filter to overcome the artifacts caused by the moving background.

Obtaining a minimum required image quality in X-ray angiography is highly challenging since different types of interventions may require different image qualities. Our results show the potential of our method for reducing the fluoroscopy time for pediatric cardiac interventions. In this work we only focused on the pulse rate and fluoroscopy time reduction since our dataset was retrospective. Other dose indicators such as cumulative air kerma should be considered along with fluoroscopy time in our future work.

Significant efforts have been invested in improving the new generation of X-ray devices, given the importance of radiation dose reduction not only for pediatric patients with high potential risks of cancer but also for adult patients, cardiologists and medical staff[26 27 28]. This study can thus pave the way for the next generation of X-ray imaging devices, as it allows to optimize the induced radiation dose for patients and staff.

Future work will consider incorporating the heart cycle information using the ECG signal for more accurate motion estimation. Other model-based or hybrid approaches can be investigated to improve the accuracy of motion prediction. Additionally, video super-resolution methods can be included in the content predictor to improve the image quality of predictions.

V. Conclusion

This work presents a novel radiation dose management approach for pediatric interventional cardiology using a learning-based video frame prediction. Such a prediction can reduce the amount of accumulated radiation dose for patients and staff by exposing them to fewer pulses while preserving the frame rate and the image quality.

Acknowledgment

The authors would like to thank Sainte-Justine Hospital Cath lab technicians and Duke university (CVIT) for their time and valuable advice. This work was supported in part by the NSERC Discovery grant and by the National Institutes of Health biomedical resource grant P41-EB028744. The Quadro RTX6000 used for this research was donated by the NVIDIA Corporation.

References

[1] Y. Liu, S. Chen, L. Zühlke, G. C. Black, M.-k. Choy, N. Li, and B. D. Keavney, Global birth prevalence of congenital heart defects 1970–2017: updated systematic review and meta-analysis of 260 studies, International journal of epidemiology **48**, 455–463 (2019).

[2] N. A. Haas, C. M. Happel, M. Mauti, C. Sahyoun, L. Z. Tebart, D. Kececioglu, and K. T. Laser, Substantial radiation reduction in pediatric and adult congenital heart disease interventions with a novel X-ray imaging technology, IJC Heart & Vasculature **6**, 101–109 (2015).

[3] A. J. Gislason-Lee, A. Kumcu, S. M. Kengyelics, D. S. Brettle, L. A. Treadgold, M. Sivananthan, and A. G. Davies, How much image noise can be added in cardiac X-ray imaging without loss in perceived image quality?, Journal of Electronic Imaging **24**, 051006 (2015).

[4] P. Hunold, F. M. Vogt, A. Schmermund, J. F. Debatin, G. Kerkhoff, T. Budde, R. Erbel, K. Ewen, and J. Barkhausen, Radiation exposure during cardiac CT: effective doses at multi–detector row CT and electron-beam CT, Radiology **226**, 145–152 (2003).

[5] K. Chida, T. Ohno, S. Kakizaki, M. Takegawa, H. Yuuki, M. Nakada, S. Takahashi, and M. Zuguchi, Radiation dose to the pediatric cardiac catheterization and intervention patient, American Journal of Roentgenology **195**, 1175–1179 (2010).

[6] L. T. Dauer, Radiation Dose Management for Fluoroscopically-Guided Interventional Procedures, (2011).

[7] K. Yamagata, B. Aldhoon, and J. Kautzner, Reduction of fluoroscopy time and radiation dosage during catheter ablation for atrial fibrillation, Arrhythmia & electrophysiology review **5**, 144 (2016).

[8] J. W. Hirshfeld et al., ACCF/AHA/HRS/SCAI clinical competence statement on physician knowledge to optimize patient safety and image quality in fluoroscopically guided invasive cardiovascular procedures: a report of the American College of Cardiology Foundation/American Heart Association/American College of Physicians Task Force on Clinical Competence and Training, Journal of the American College of Cardiology **44**, 2259–2282 (2004).

[9] S. Balter, Fluoroscopic frame rates: not only dose, American Journal of Roentgenology **203**, W234–W236 (2014).

[10] C. T. Pyne, G. Gadey, C. Jeon, T. Piemonte, S. Waxman, and F. Resnic, Effect of reduction of the pulse rates of fluoroscopy and CINE-acquisition on X-ray dose and angiographic image quality during invasive cardiovascular procedures, Circulation: Cardiovascular Interventions **7**, 441–446 (2014).

[11] R. Aufrichtig, P. Xue, C. W. Thomas, G. C. Gilmore, and D. L. Wilson, Perceptual comparison of pulsed and continuous fluoroscopy, Medical physics **21**, 245–256 (1994).

[12] D. L. Wilson, P. Xue, and R. Aufrichtig, Perception of fluoroscopy last-image hold, Medical physics **21**, 1875–1883 (1994).

[13] F. Azizmohammadi, R. Martin, J. Miro, and L. Duong, Model-free cardiorespiratory motion prediction from X-ray angiography sequence with LSTM network, in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7014–7018, IEEE, (2019).

[14] H. Fang, H. Li, S. Song, K. Pang, D. Ai, J. Fan, H. Song, Y. Yu, and J. Yang, Motion-flow-guided recurrent network for respiratory signal estimation of X-ray angiographic image sequences, Physics in Medicine & Biology (2020).

[15] Q. Lyu, H. Shan, Y. Xie, D. Li, and G. Wang, Cine Cardiac MRI Motion Artifact Reduction Using a Recurrent Neural Network, arXiv preprint arXiv:2006.12700 (2020).

[16] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, Generative Adversarial Networks: An Overview, IEEE Signal Processing Magazine **35**, 53–65 (2018).

[17] Z. Hu and J. T. Wang, Generative Adversarial Networks for Video Prediction with Action Control, in *International Joint Conference on Artificial Intelligence*, pages 87–105, Springer, (2019).

[18] M. Mathieu, C. Couprie, and Y. LeCun, Deep multi-scale video prediction beyond mean square error, arXiv preprint arXiv:1511.05440 (2015).

[19] W. P. Segars, G. Sturgeon, S. Mendonca, J. Grimes, and B. M. Tsui, 4D XCAT phantom for multimodality imaging research, Medical physics **37**, 4902–4915 (2010).

[20] T. Kobayashi and J. W. Hirshfeld Jr, Radiation exposure in cardiac catheterization: operator behavior matters, (2017).

[21] M. Mahesh, Fluoroscopy: patient radiation exposure issues, Radiographics **21**, 1033–1045 (2001).

[22] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, Decomposing motion and content for natural video sequence prediction, arXiv preprint arXiv:1706.08033 (2017).

[23] J.-T. Hsieh, B. Liu, D.-A. Huang, L. F. Fei-Fei, and J. C. Niebles, Learning to decompose and disentangle representations for video prediction, in *Advances in Neural Information Processing Systems*, pages 517–526, (2018).

[24] S. Tulyakov, M. Y. Liu, X. Yang, and J. Kautz, Mocogan: Decomposing motion and content for video generation, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, (2018).

[25] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, Multiscale vessel enhancement filtering, in *International conference on medical image computing and computer-assisted intervention*, pages 130–137, Springer, (1998).

[26] A. J. Gislason-Lee, C. Keeble, C. J. Malkin, D. Egleston, J. Bexon, S. M. Kengyelics, D. Blackman, and A. G. Davies, Impact of latest generation cardiac interventional X-ray equipment on patient image quality and radiation dose for trans-catheter aortic valve implantations, The British Journal of Radiology **89**, 20160269 (2016).

[27] A. J. Gislason-Lee, C. Keeble, D. Egleston, J. Bexon, S. M. Kengyelics, and A. G. Davies, Comprehensive assessment of patient image quality and radiation dose in latest generation cardiac x-ray equipment for percutaneous coronary interventions, Journal of Medical Imaging **4**, 025501 (2017).

[28] A. H. McNeice, M. Brooks, C. G. Hanratty, M. Stevenson, J. C. Spratt, and S. J. Walsh, A retrospective study of radiation dose measurements comparing different cath lab X-ray systems in a sample population of patients undergoing percutaneous coronary intervention for chronic total occlusions, Catheterization and Cardiovascular Interventions **92**, E254–E261 (2018).

List of Figures

Table 1: Average similarity measurements of the predicted images over the testing data on three predicted images for simulated dataset.

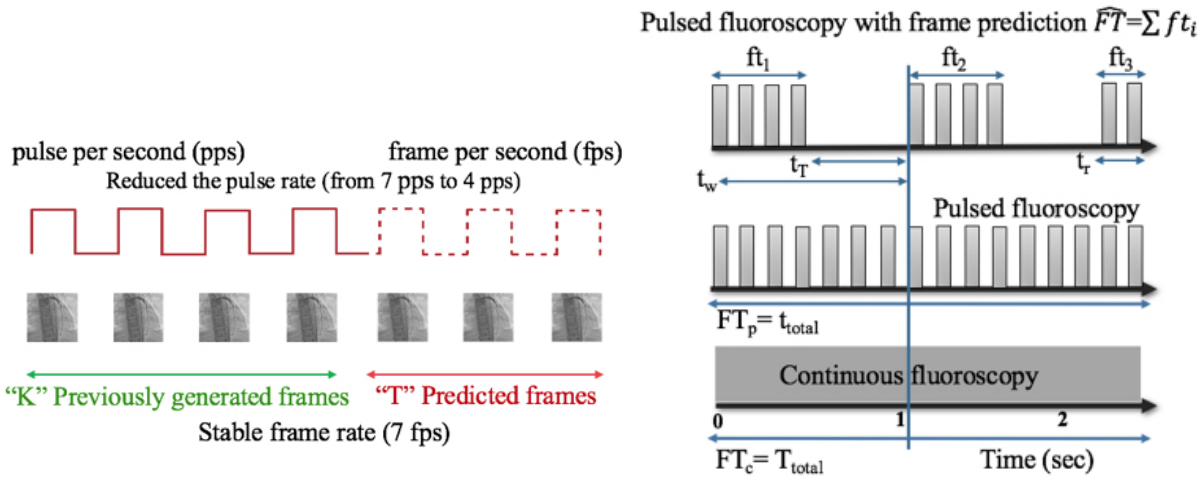
Frame	Vss PSNR	Original MCnet PSNR	Vss SSIM	Original MCnet SSIM
Simulated data K=7, T=1,2,3				
Frame 1	28.28	27.98	0.94	0.89
Frame 2	25.47	24.85	0.92	0.85
Frame 3	23.90	23.01	0.88	0.82
Simulated data K=10, T=1,2,3				
Frame 1	29.13	28.82	0.97	0.86
Frame 2	27.65	25.10	0.93	0.83
Frame 3	24.14	23.12	0.87	0.81

Table 2: Average similarity measurements of the predicted images over the testing data on three predicted images for patient dataset.

Frame	Vss PSNR	Original MCnet PSNR	Vss SSIM	Original MCnet SSIM
Patient data K=7, T=1,2,3				
Frame 1	27.10	26.75	0.79	0.80
Frame 2	24.42	23.59	0.68	0.70
Frame 3	23.10	21.54	0.61	0.61
Patient data K=10, T=1,2,3				
Frame 1	27.97	26.80	0.82	0.78
Frame 2	25.65	24.62	0.74	0.69
Frame 3	24.14	23.32	0.68	0.63

Table 3: Euclidean distance between the centrelines of arteries in the predicted frames and ground truth for the frame prediction and gap-filling.

# Frame	Euclidean distance (mm)					
	Frame prediction			Gap-filling		
	Mean	Max	SD	Mean	Max	SD
Frame 1	0.28 mm	0.76 mm	(+/-) 0.19 mm	0.33 mm	0.79 mm	(+/-) 0.22 mm
Frame 2	0.30 mm	0.78 mm	(+/-) 0.20 mm	0.39 mm	0.85 mm	(+/-) 0.31 mm
Frame 3	0.32 mm	0.84 mm	(+/-) 0.21 mm	0.51 mm	0.93 mm	(+/-) 0.35 mm



(a) (b)

Figure 1: (a) The sequence at 7 fps frame rate is acquired partially with exposed pulses and partially with predictions such that the pulse rate gets reduced while the frame rate remains constant ($K = 4$ and $T = 3$). (b) An example of three different fluoroscopy techniques. Less fluoroscopy time is required for pulsed discrete fluoroscopy by pausing the radiation beam after K acquired images for a prediction time t_T in each time window t_w compared to other methods ($\hat{FT} < FT_p < FT_c$).

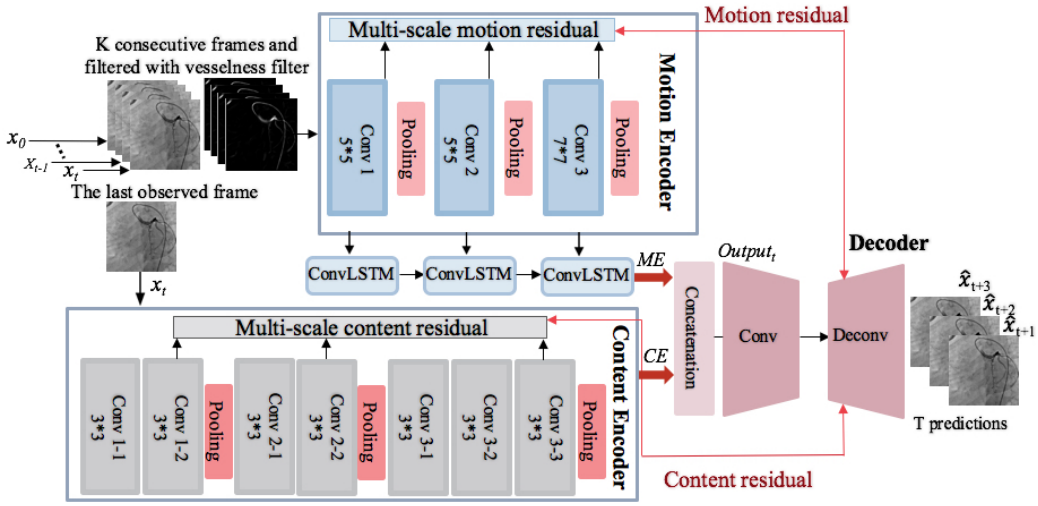


Figure 2: The motion-content model structure. Two encoders extract the motion and content features separately (ME and CE). The input for the motion encoder is a sub-sequence of previously acquired and visited frames filtered by the vesselness filter. The input for the content encoder is the last visited frame. The outputs of these two encoders are concatenated to be decoded as a sub-sequence of predictions. The motion and content residuals are added to avoid information loss.

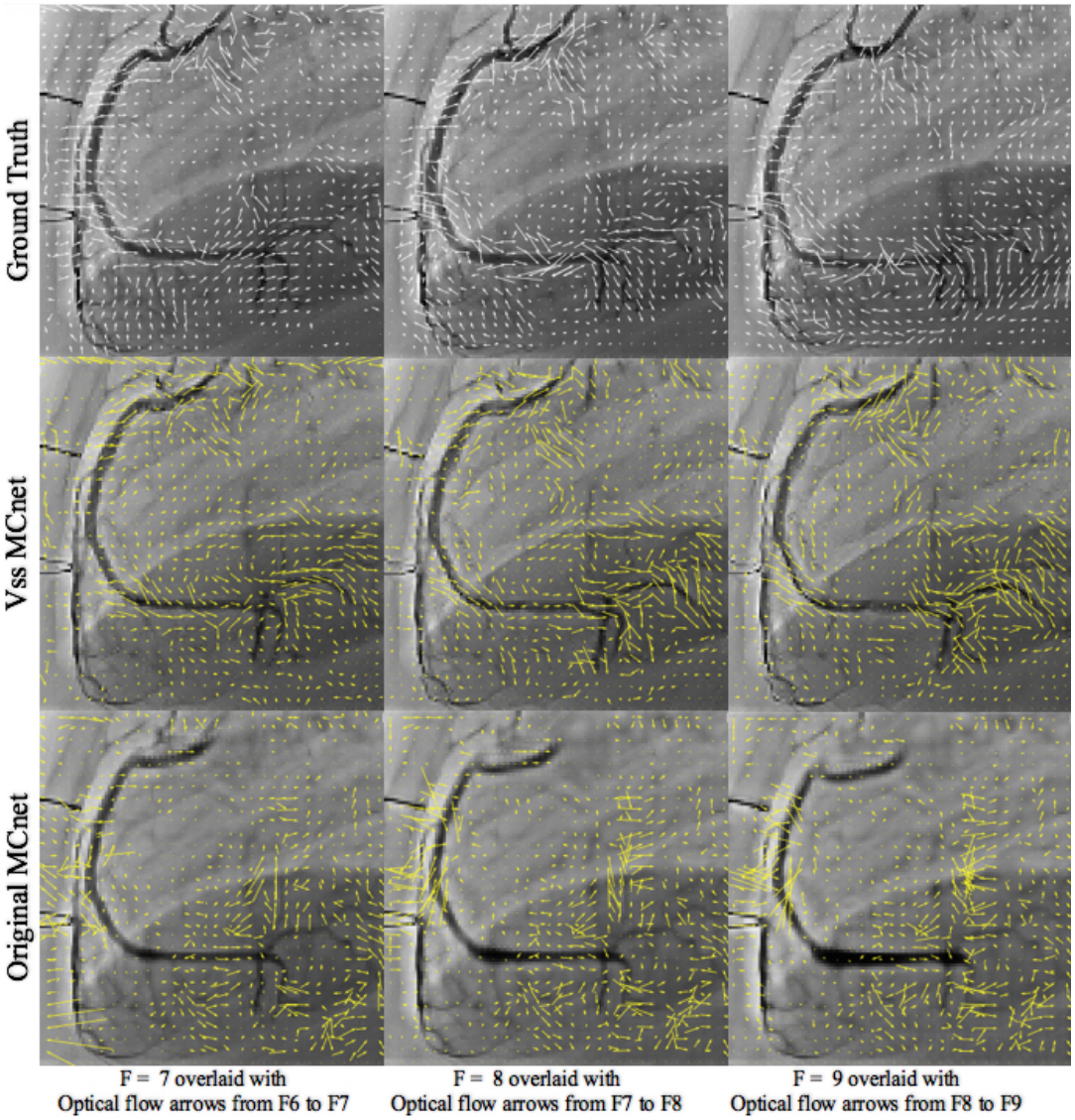


Figure 3: Optical flow estimated motion fields of the ground truth sequence (top-white arrows) and the generated frames with vesselness-based MCnet on the second row and original MCnet on third row (yellow arrows). The optical flow fields are overlaid to the predicted frames F7,F8,F9.

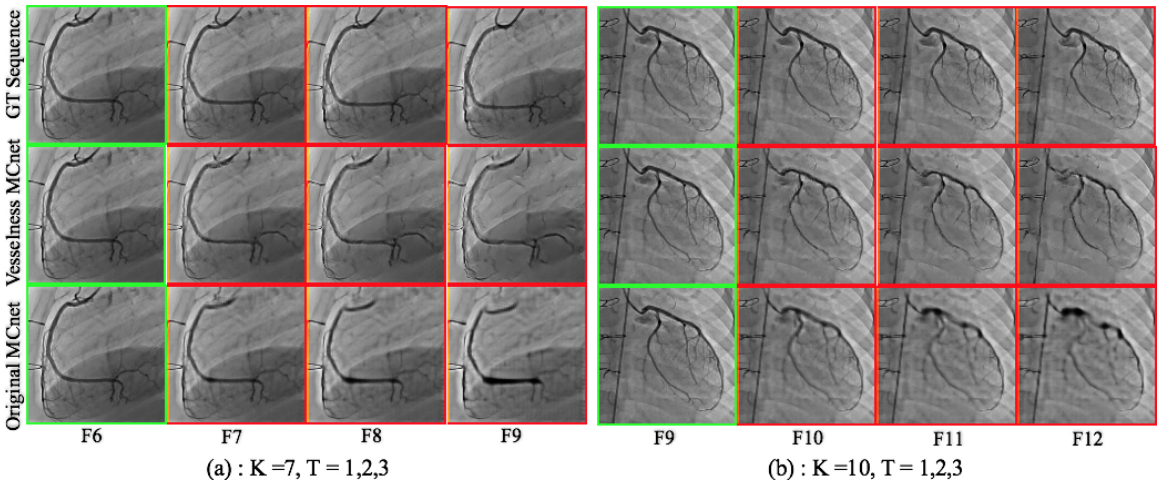


Figure 4: The first row shows the ground truth sequence. The second and third rows show the results of vesselness-based MCnet and original MCnet, respectively. The predicted images are identified with a red outline and the last visited frame with a green outline.

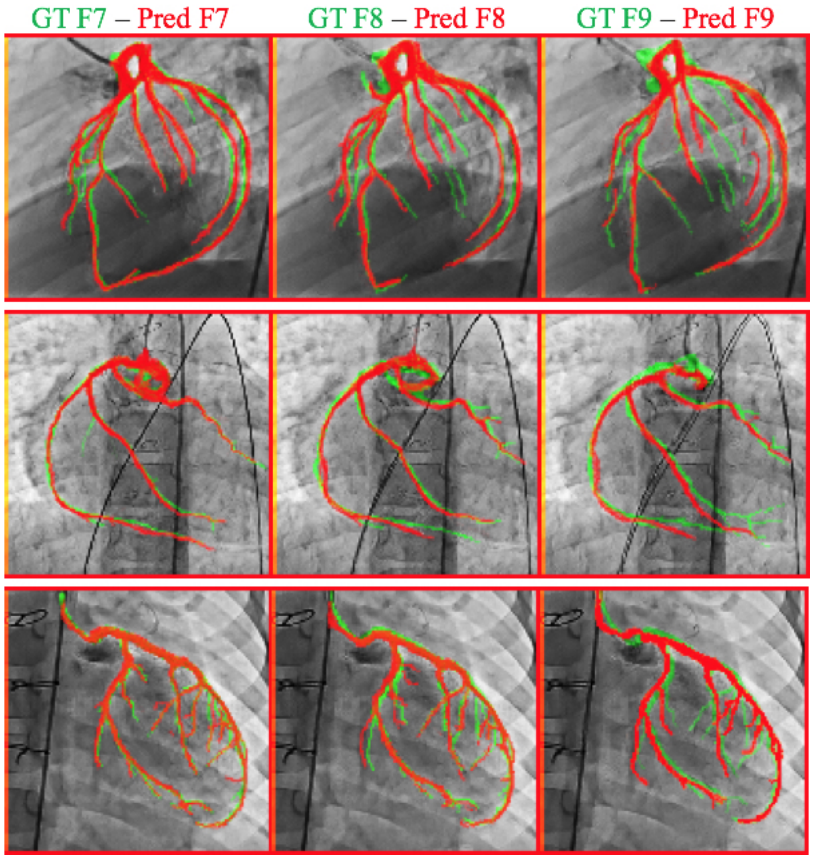


Figure 5: An overlay of the manual segmentation masks for the ground truth in green and predicted sequences in red.