MR to X-ray Projection Image Synthesis

Bernhard Stimpel, Christopher Syben, Tobias Würfl, Katrin Mentl, Arnd Dörfler, and Andreas Maier

***Abstract*—Hybrid imaging promises large potential in medical imaging applications. To fully utilize the possibilities of corresponding information from different modalities, the information must be transferable between the domains. In radiation therapy planning, existing methods make use of reconstructed 3D magnetic resonance imaging data to synthesize corresponding X-ray attenuation maps. In contrast, for fluoroscopic procedures only line integral data, i.e., 2D projection images, are present. The question arises which approaches could potentially be used for this MR to X-ray projection image-to-image translation. We examine three network architectures and two loss-functions regarding their suitability as generator networks for this task. All generators proved to yield suitable results for this task. A cascaded refinement network paired with a perceptual-loss function achieved the best qualitative results in our evaluation. The perceptual-loss showed to be able to preserve most of the high-frequency details in the projection images and, thus, is recommended for the underlying task and similar problems. The abstract goes here.**

***Index Terms*—Medical image synthesis, multi-modality fusion, machine learning, Fluoroscopy**

I. Introduction

Promising concepts on how a combined magnetic resonance (MR) and computed tomography (CT) imaging device may look like were proposed in the past. Wang et al. [1] published a top-level design of an MR-CT scanner consisting of two superconducting electromagnets surrounding multiple, rotatable X-ray sources. The desired application for their model is combined image reconstruction for plaque characterization. In contrast, [2] focused on the interventional applicability of a hybrid MR-X-ray system and showed the great potential of this application. Assuming an imaging device that is capable of acquiring corresponding X-ray and MR projection images simultaneously, or at least consecutively in the same state of motion, the combined information would be highly useful for fluoroscopic procedures. On the one hand, overlay strategies of both modalities in their respective form could be used to simultaneously visualize soft- and dense-tissue or -material. On the other hand, the information of one modality could be transferred to the domain of its counterpart. This information could then be used for further processing and image enhancement. A possible application would be to exploit the high signal-to-noise ratio of MR imaging, especially in soft-tissue regions, to apply denoising methods on the corresponding X-ray images. Considering that the noise level in X-ray Fluoroscopy is directly related to the applied radiation dose, a higher tolerance for noise could lead to reduction of harmful patient radiation exposure. Furthermore, it allows for investigations in the field of super-resolution. Most of the mentioned applications would require corresponding images in the same domain. The acquisition of projection images that match the typical projective distortion directly from the MR is possible, as shown by [3], [4]. To allow for further downstream processing, a possibility to transfer the information between the projection images in the distinct domains would be useful. Similar methods are already used in radiation therapy planning, where attenuation maps are estimated from pseudo-CT scans that are synthesized from corresponding MR data [5], [6], [7]. However, all these methods are based on 3D tomographic image data. In contrast, for fluoroscopic procedures this transfer between the domains must be performed based on line integral data, i.e., 2D projection images, and not on reconstructed images. Motivated by its possible applications and inspired by existing methods from radiation therapy and natural image synthesis, we investigate different deep learning-based methods for X-ray projection image synthesis from MR projections.

This work has been supported by the project P3-Stroke, an EIT Health innovation project. EIT Health is supported by EIT, a body of the European Union.

B. Stimpel, C. Syben, T. Würfl, K. Mentl, and A. Maier are with Friedrich-Alexander-Universität Erlangen-Nürnberg, Pattern Recognition Lab, Erlangen, Germany.

B. Stimpel, C.Syben, and A. Dörfler are also with Friedrich-Alexander-Universität Erlangen-Nürnberg, Department of Neuroradiology, Erlangen, Germany.

II. Methods

Convolutional neural networks have shown great results in natural and medical image synthesis [6], [8]. Based on this, three different generator network architectures are used in the underlying work with the goal to generate X-ray projections $\boldsymbol{G}$ from input MR projection images $\boldsymbol{I}$. Training and evaluation are done using corresponding MR and label X-ray projections $\boldsymbol{L}$. All models have been adapted to our specific application. An overview of the investigated network architectures is given in Figure 1. Furthermore, we examined the impact of two different loss-functions on the generated results.

A. Model Architecture

Convolutional auto-encoders are a popular choice for generator networks in image synthesis. In general, an auto-encoder consists of an encoder and a decoder path. In the encoder path the image's resolution is decreased and the filter dimension is increased. The subsequent decoder path reverts this process to reach the initial resolution and dimension again. Enhancing the encoder-decoder structure with skip-connections between corresponding resolution levels has proven to be beneficial regarding the conservation of spatial information lost during down-sampling. Our first network model is close to the well-known "U-net" introduced by Ronneberger et al. [9]. Instead of maximum pooling layers we use strided convolution with stride two for up- and down-sampling. In addition to the

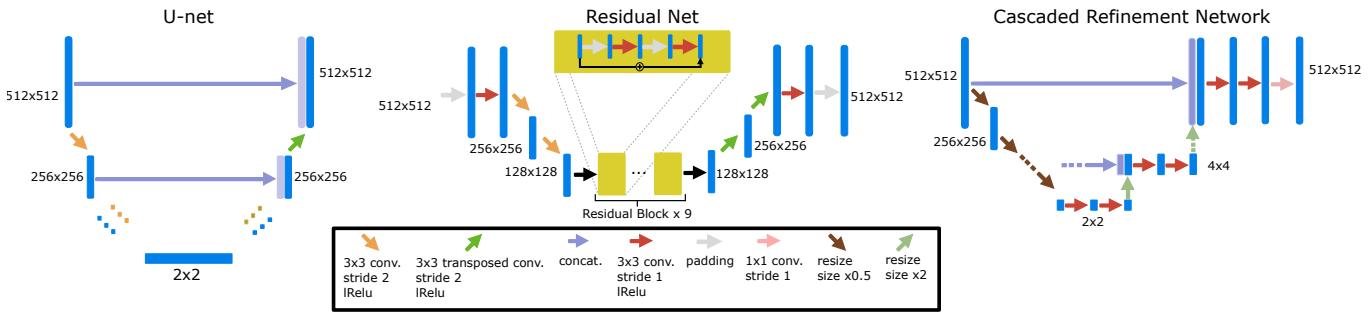


Fig. 1: Schematic architectures of the different generator networks.

architecture presented in Figure 1, the first three layers of the synthesis path use dropout with a keep probability of 50 percent.

The second generator network is a deep residual network (ResNet) [10] which was initially proposed for image recognition. The key component of this approach are residual connections that allow for more robust training of deeper networks than before. Besides the original application, this network architecture proved to yield good result in generative tasks. We use the model proposed by [11] for style transfer to generate our estimated X-ray projections. Deviating from their proposal, we add nine residual blocks instead of the originally proposed five.

Finally, a cascaded refinement network (CRN) is used as image generator. This model was recently proposed by Chen et al. [12] and yielded good results on natural image synthesis from a semantic layout. In contrast to many currently proposed approaches, their model does not use adversarial training but relies on a single feedforward network. The semantic layout as input is replaced by MR projection images in our case. The network consists of multiple refinement modules that work in a multi-scale strategy from coarse to fine as presented in Figure 1. The full model is built from 8 single refinement modules and the final 1 x 1 convolution layer maps the output to a single channel image. A major difference to the first two network architectures is that Chen et al. relinquished convolutional layers in the down-scaling path and, instead, only use resizing operations. Input information from higher resolution scales is solely incorporated using concatenation. By this, additional model capacity can be used for the subsequent up-scaling path.

B. Objective Functions

The choice of the objective function is a key aspect in every machine learning application. Multiple functions have been used for the task of image-to-image translation and image synthesis in the past. We picked two different loss-functions to compare them in our approach. Since a one-to-one correspondence is given by the matching image pairs, a simple but suitable loss function for image generation tasks is the ℓ_1-norm [13]. Pixel-wise comparison of the generated and label image intensities via the ℓ_1-loss function can be done by calculating

$$\mathrm{E}_{\ell_1}(\boldsymbol{L}, \boldsymbol{G}) = \sum_{i}^{N} |\boldsymbol{L}(i) - \boldsymbol{G}(i)| \ , \tag{1}$$

where i denotes one image pixel, $i \in N$, and N is the number of all pixel in one image.

A second loss function that was recently proposed for natural image synthesis without corresponding image pairs is the perceptual-loss [11]. The perceptual-loss does not calculate the error between the estimated and real intensity values. Instead, the generated and the label image are fed into a pre-trained image classification network that we will refer to as evaluation network in the following. While the resulting classification scores are not of interest, the raw feature activations between the different input images are compared. The underlying theory is that similarly looking images activate the same units in the image classification network, i.e. the higher the accordance between both feature activations the more similar the generated and label image are. The loss function can be written as

$$\mathrm{E}_{\mathrm{p}}(\boldsymbol{L}, \boldsymbol{G}) = \sum_{k}^{K} \left(\boldsymbol{V}_k(\boldsymbol{L}) - \boldsymbol{V}_k(\boldsymbol{G}) \right) \ , \tag{2}$$

where $\boldsymbol{V}_k(\boldsymbol{L})$ and $\boldsymbol{V}_k(\boldsymbol{G})$ is the feature activation map of the evaluation network for the label image $\boldsymbol{L}$ and the generated image $\boldsymbol{G}$ at the current layer k, $k \in K$. In this approach, the perceptual-loss is computed on the VGG-19 network [14] which was pre-trained on the ImageNet data set [15].

All generators are trained with an ADAM optimizer [16] and a learning rate of 0.004 for 100 epochs.

III. Experiments

Experiments were conducted using data of a realistic MR and X-ray sensitive phantom of the human head. Data was acquired on a 1.5 T Aera MR and a Axiom-Artis C-arm CT scanner (Siemens Healthcare GmbH, Forchheim, Germany). An ultra-short echo time sequence was used for the MRI scans. The reconstructed images' resolution is 320 x 320 x 250 with a spacing of 0.93 x 0.93 x 0.93 mm^3. The X-ray scans of the same phantom exhibit a voxel size of 0.48 x 0.48 x 0.48 mm^3 and a resolution of 512 x 512 x 399. Image registration of the corresponding scans was performed using elastix. The input (MR) and label (CT) images were generated by forward projecting the registered stack from various angulations using the CONRAD framework [17]. In this manner, 3200 different projection image pairs of both modalities were created and randomly divided into 3000 training and 200 testing images.

	MAE	SSIM	PSNR
U-net - p-loss	0.083	0.891	26.994
ResNet - p-loss	0.077	0.924	27.675
CRN - p-loss	0.071	0.931	28.353
U-net - l1-loss	0.068	0.917	28.506
ResNet - l1-loss	0.058	0.938	30.067
CRN - l1-loss	0.084	0.920	27.097

TABLE I: Quantitative results of the different network architectures and loss functions

The evaluation of the output can be done by calculating the deviation of the generated X-ray $\boldsymbol{G}$ from the real X-ray images $\boldsymbol{L}$. The mean squared error (MSE) can be used to this end. It is computed as

$$\text{MSE}(\boldsymbol{L}, \boldsymbol{G}) = \frac{1}{N} \sum_{i}^{N} \left\| \boldsymbol{L}(i) - \boldsymbol{G}(i) \right\|_2^2 . \tag{3}$$

Yet, not only the absolute difference of estimated values is of interest in projection image synthesis. The generated projection images must also correspond to each other from a visual point of view, which cannot be determined entirely be pixel-wise comparison of the image pairs. To this end, the structural similarity (SSIM) index [18], a perception-based metric, is computed. Assuming two patches $\boldsymbol{g}$ and $\boldsymbol{l}$ of the generated and label image. The SSIM is then computed as

$$\text{SSIM}(\boldsymbol{g}, \boldsymbol{l}) = \frac{(2\mu_{\boldsymbol{g}}\mu_{\boldsymbol{l}} + c_1)(2\sigma_{\boldsymbol{gl}} + c_2)}{(\mu_{\boldsymbol{g}}^2 + \mu_{\boldsymbol{l}}^2 + c_1)(\sigma_{\boldsymbol{g}}^2 + \sigma_{\boldsymbol{l}}^2 + c_2)} , \tag{4}$$

where μ is the mean, σ^2 the variance, and σ the covariance. To avoid instabilities, the constants c_1 and c_2 are introduced that are defined as $c_i = (K_i \mathcal{L})^2$, $i \in \{1, 2\}$, with $\mathcal{L}$ being the dynamic range of the intensity values and $K_1 = 0.01$ and $K_2 = 0.03$. Computing Equation 4 for all pairs of patches $\boldsymbol{g}$ and $\boldsymbol{l}$ yields the final SSIM measure for the whole image.

The third evaluation metric that is computed is the peak signal-to-noise ratio (PSNR). The PSNR measures the ratio between the highest intensity value and the occuring noise and is often applied to measure image quality, especially regarding reconstruction and compression loss. It is computed by

$$\text{PSNR}(\boldsymbol{L}, \boldsymbol{G}) = 20 \log_{10} \frac{\max(\boldsymbol{G})}{\text{MSE}(\boldsymbol{L}, \boldsymbol{G})} . \tag{5}$$

In the subsequent chapter results for all metrics will be presented. To present comparable absolute numbers, all images were scaled from -1 to 1 prior to the error metric calculations.

IV. Results and Discussion

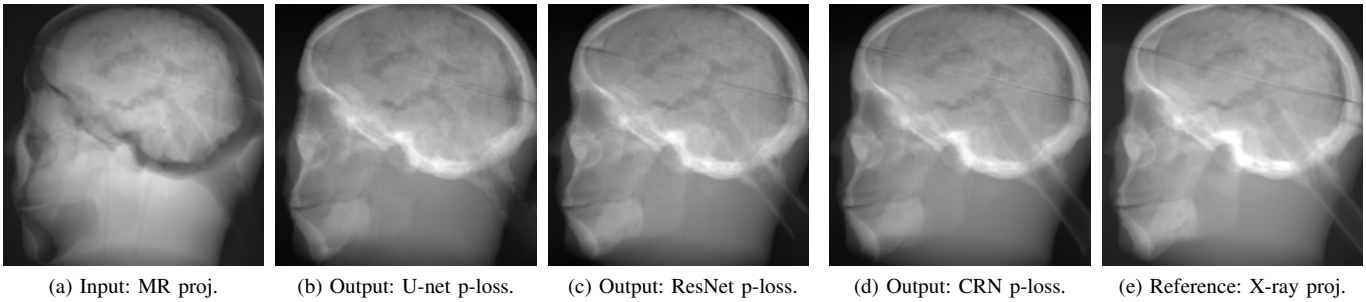
The quantitative and qualitative results of the proposed experiments are presented in Table I and Figure 2. By examining these it can be observed that the differences in the calculated MSE of all network architectures and incorporated loss functions are only small. The best results in terms of pixel-wise deviation could be achieved with the ResNet architecture combined with the ℓ_1-loss function. This network achieves a deviation from the reference of only 0.058, i.e., 2.4 percent. Also the results of the U-net and CRN networks are still good with deviations of 2.6 and 2.9 percent. Similarly small variation can be observed in the structured similarity measure. The ResNet and CRN exhibit approximately equal quality with SSIM measures of 0.938 and 0.920 for the ℓ_1-loss and 0.924 and 0.931 for the perceptual-loss, respectively. The results generated with the U-net are slightly worse. The highest peak signal-to-noise ratio is achieved by the ResNet (ℓ_1-loss), followed by the U-net (ℓ_1-loss) and CRN (p-loss). It is noteworthy that the ResNet and U-net both achieve the highest results in all error metrics using the ℓ_1-loss while the opposite is the case for the CRN which works best with the perceptual-loss function.

Overall, the perceptual-loss achieves competitive and in some cases even better results than the ℓ_1-loss when comparing the pixel-wise error metrics. For example, the cascaded refinement network's MSE is 0.013 smaller for the perceptual- than for the ℓ_1-loss. This might be suspicious at first sight, considering that the ℓ_1-loss purely optimizes for this pixel-wise error in the training process while the perceptual-loss compares the raw feature activations of the evaluation network. Contrarily, this behavior cannot be observed for the U-net and ResNet. The results produced with the ℓ_1-loss achieve higher values for all error measures for these networks. An explanation for this obervation is that the intensity values of the input image still cause an impact on the respective layers output in the evaluation network when computing the perceptual-loss. Consequently, these differences also transition to the computed loss value for all feature layers. Even though the perceptual-loss incorporates the raw intensity values, it is not guaranteed that the scaling of these is conserved in this process. By this, the relative changes can be similar, whereas the absolute range of values changes and, correspondingly, also the pixel-wise error metrics.

Another observation is that the perceptual-loss is able to conserve high-frequency details in the image. The fine line in the projection images that forms a circle around the cranium is visible in the input (Figures 2a & 2f), as well as in the label images (Figures 2e & 2j), and also in the images generated with the perceptual-loss function (Figures 2b, 2c, and 2d). In contrast, all generators "loose" this line when the ℓ1-loss is applied (Figures 2g, 2h, and 2i). This effect is also qualitatively observable in other parts of the images. Despite achieving equal or better results regarding the error metrics, the generally less sharp look of the results generated with the ℓ1-loss function is apparent. This behavior is in accordance with previous observations that concluded that an perceptual-loss leads to sharper images than a comparable ℓ1-loss [19]. Considering the common applications of X-ray Fluoroscopy, e.g., interventional guidance for stents and similar devices, high spatial resolution is a key requirement. Utilizing a loss function that is able to preserve high-frequency details in the images is desirable to this end. The perceptual-loss appears to be suited for this task as presented in our evaluation.

V. Conclusion

We showed the feasibility of image-to-image translation from MR projection images to corresponding X-ray projections. Three generator networks and two different loss



(a) Input: MR proj. (b) Output: U-net p-loss. (c) Output: ResNet p-loss. (d) Output: CRN p-loss. (e) Reference: X-ray proj.

(f) Input: MR proj. (g) Output: U-net ℓ_1-loss. (h) Output: ResNet ℓ_1-loss. (i) Output: CRN ℓ_1-loss. (j) Reference: X-ray proj.

Fig. 2: Results of the projection synthesis. Top row: Results generated with the perceptual-loss function. Bottom row: Results generated with the ℓ_1-loss function.

functions were implemented and evaluated to this end. All examined network architectures achieved good results on the proposed task. When comparing the generated projection images of all networks it became apparent that the loss function has a greater impact on the images' quality than the actual architectures of the network. The perceptual-loss proved to be able to conserve even small high-frequency details in the course of the image-to-image transfer. Because high-spatial resolution is desired in most fluoroscopic procedures, we recommend using this perceptual-loss function for the underlying task. The best quantitative and qualitative results with this loss function could be achieved by a cascaded refinement model in this work. The high-quality of the generated projection images unveils large potential regarding the applicability to multi-modal denoising, super-resolution, and more. As a next step, we plan to transfer this approach to real patient data. Additionally, the effect of combining multiple different MR acquisition protocols and weighting schemes will be investigated.

REFERENCES

[1] G. Wang, F. Liu, F. Liu, G. Cao, H. Gao, and M. W. Vannier, "Top-Level Design of the first CT-MRI scanner," in *Proc. 12th Fully 3D Meet.*, 2013, pp. 5–8.

[2] R. Fahrig, K. Butts, J. A. Rowlands, R. Saunders, J. Stanton, G. M. Stevens, B. L. Daniel, Z. Wen, D. L. Ergun, and N. J. Pelc, "A truly hybrid interventional MR/x-ray system: Feasibility demonstration," *J. Magn. Reson. Im.*, vol. 13, no. 2, pp. 294–300, 2001.

[3] S. Napel, S. Dunne, and B. K. Rutt, "Fast Fourier projection for MR angiography," *Magn. Reson. Med.*, vol. 19, no. 2, pp. 393–405, 1991.

[4] C. Syben, B. Stimpel, M. Leghissa, A. Dörfler, and A. Maier, "Fan-beam Projection Image Acquisition using MRI," in *3rd Conf. Image-Guided Interv. Fokus Neuroradiol.*, M. Skalej and C. Hoeschen, Eds., 2017, pp. 14–15.

[5] B. K. Navalpakkam, H. Braun, T. Kuwert, and H. H. Quick, "Magnetic ResonanceBased Attenuation Correction for PET/MR Hybrid Imaging Using Continuous Valued Attenuation Maps," *Invest. Radiol.*, vol. 48, no. 5, pp. 323–332, 2013.

[6] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical Image Synthesis with Context-Aware Generative Adversarial Networks," *Med. Image Comput. Comput. Interv. MICCAI 2017*, pp. 417–425, 2017.

[7] J. M. Wolterink, A. M. Dinkla, M. H. F. Savenije, P. R. Seevinck, C. A. T. van den Berg, and I. Išgum, "Deep MR to CT Synthesis Using Unpaired Data," in *Int. Work. Simul. Synth. Med. Imaging*, 2017, pp. 14–23.

[8] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," *IEEE Conf. Comput. Vis. pattern Recognit.*, pp. 2414–2423, 2016.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Med. Image Comput. Comput. Interv. MICCAI 2015*, pp. 234–241, 2015.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.

[11] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," *arXiv:1603.08155*, 2016.

[12] Q. Chen and V. Koltun, "Photographic Image Synthesis with Cascaded Refinement Networks," in *Int. Conf. Comput. Vis.*, 2017.

[13] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss Functions for Image Restoration with Neural Networks," *IEEE Trans. Comput. IMAGING*, vol. 3, no. 1, 2017.

[14] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556*, 2015.

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2014.

[16] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Int. Conf. Learn. Represent.*, 2015.

[17] A. Maier, H. G. Hofmann, M. Berger, P. Fischer, C. Schwemmer, H. Wu, K. Müller, J. Hornegger, J.-H. Choi, C. Riess, A. Keil, and R. Fahrig, "CONRAD - A software framework for cone-beam imaging in radiology," *Med. Phys.*, vol. 40, no. 11, pp. 111914, 2013.

[18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[19] A. Dosovitskiy and T. Brox, "Generating Images with Perceptual Similarity Metrics based on Deep Networks," *Adv. Neural Inf. Process. Syst. 29 (NIPS 2016)*, pp. 658–666, 2016.