Context-Aware, Reference-Free Local Motion Metric for CBCT Deformable Motion Compensation

Heyuan Huang[a], Jeffrey H. Siewerdsen[a,b], Wojciech Zbijewski[a], Clifford R. Weiss[b],
Mathias Unberath[c], Alejandro Sisniega*[a]
[a] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD; [b]Department of Radiology, Johns Hopkins University, Baltimore, MD; [c]Department of Computer Science, Johns Hopkins University, Baltimore, MD.
*email: asisniega@jhu.edu

ABSTRACT

Deformable motion is one of the main challenges to image quality in interventional cone beam CT (CBCT). Autofocus methods have been successfully applied for deformable motion compensation in CBCT, using multi-region joint optimization approaches that leverage the moderately smooth spatial variation motion of the deformable motion field with a local neighborhood. However, conventional autofocus metrics enforce images featuring sharp image-appearance, but do not guarantee the preservation of anatomical structures. Our previous work (DL-VIF) showed that deep convolutional neural networks (CNNs) can reproduce metrics of structural similarity (visual information fidelity - VIF), removing the need for a matched motion-free reference, and providing quantification of motion degradation and structural integrity. Application of DL-VIF within local neighborhoods is challenged by the large variability of local image content across a CBCT volume and requires global context information for successful evaluation of motion effects. In this work, we propose a novel deep autofocus metric, based on a context-aware, multi-resolution, deep CNN design. In addition to the inclusion of contextual information, the resulting metric generates a voxel-wise distribution of reference-free VIF values. The new metric, denoted CADL-VIF, was trained on simulated CBCT abdomen scans with deformable motion at random locations and with amplitude up to 30 mm. The CADL-VIF achieved good correlation with the ground truth VIF map across all test cases with $R^2 = 0.843$ and slope = 0.941. When integrated into a multi-ROI deformable motion compensation method, CADL-VIF consistently reduced motion artifacts, yielding an average increase in SSIM of 0.129 in regions with severe motion and 0.113 in regions with mild motion. This work demonstrated the capability of CADL-VIF to recognize anatomical structures and penalize unrealistic images, which is a key step in developing reliable autofocus for complex deformable motion compensation in CBCT.

Keywords: Interventional CBCT, Motion Compensation, Deformable Motion, Convolutional Neural Network

1. INTRODUCTION

Cone-beam CT provides 3D guidance and intraprocedural imaging in interventional radiology for abdominal procedures but relatively long acquisition time makes it susceptible to patient motion from a complex combination of various periodic and aperiodic sources.

Previous work showed successful application of autofocus optimization for rigid motion compensation [1] using only the acquired CBCT data, with extension to complex deformable motion in abdominal CBCT [2]. Autofocus methods estimate a motion trajectory by minimizing an image-based metric that encourages properties associated to motion-free images (e.g, sharpness or piece-wise constancy). However, such metrics are agnostic to the underlying anatomy and might enforce solutions that satisfy the metric but feature unrealistic structural content.

Our previous work addressed such limitation via a reference-free image similarity metric (DL-VIF) with application to rigid motion compensation in neuro CBCT [3,4]. DL-VIF leveraged the potential of deep convolutional neural networks (CNNs) to extract features specific to motion image degradation and reproduce the capability of Visual Information Fidelity (VIF) [5] to quantify image degradation and structural similarity to a matched motion-free reference, but removing the need for such reference, which is usually not available in clinical settings.

DL-VIF was trained to act on images encompassing the complete head anatomy, with a moderately coarse pixel size, and to provide a global DL-VIF score aggregating contributions to VIF from all structures in the volume into a single scalar. While those assumptions are appropriate for global, rigid, motion compensation, they present various limitations in deformable motion scenarios: i) autofocus deformable motion compensation requires estimation of local VIF, to guide the compensation algorithm towards regions of large motion while ignoring static anatomy; ii) the global nature of the metric makes it susceptible to be dominated by high-contrast structures; and, iii) the coarse voxel size (~2 mm) required for training of the 3D DL-VIF CNN might be not be sufficient for capturing subtle deformation of low-contrast structures.

7th International Conference on Image Formation in X-Ray Computed Tomography,
edited by Joseph Webster Stayman, Proc. of SPIE Vol. 12304, 1230412
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2646857

A patch-based DL-VIF could provide such metric locality, at moderate volume size, but the lack of global context and the inconsistency of image contrast and structure between soft-tissue and bone regions challenges the extraction of meaningful features, resulting in degraded performance, observed in previous attempts to CNN-based deformable autofocus [6].

In this work we propose a novel, context-aware, reference-free autofocus metric, denoted CADL-VIF, that employs a context-aware deep CNN and a voxel-based local VIF definition to provide local estimations of artifacts and structural integrity. CADL-VIF was integrated into a deformable motion compensation framework and was evaluated for soft-tissue deformable motion compensation in simulated cases.

2. MATERIAL AND METHODS

2.1 CADL-VIF: A Context-Aware Local DL-VIF Design

VIF provides an estimation of the similarity between a test image (motion-corrupted in our case) and a reference image (motion-free) by quantifying the information preserved after a distortion process (motion, in this work), weighted by a

convolution channel, designed as a surrogate of the human visual system response. VIF acts on the integral of the information across the complete image, providing a scalar output. In this work, we extend the definition of VIF to incorporate localized 3D estimation of preserved information.

The resulting VIF$_L$, was based on the approach in Ref [7] and adapted to 3D volumes. The key difference between VIF$_L$ and original VIF is that when calculating the information contained in motion-corrupted (I_{MC}) and motion-free (I_{MF}) images, VIF$_L$ preserves spatial information by omitting integration across the image dimensions, summing only across different filter channels, as shown in Eqs. 1 and 2.

$$I_{MC} = \sum_{\substack{Filter\\Channels}} \log(1 + \frac{g^2 \cdot \sigma_{MF}^2}{\sigma_v^2 + \sigma_n^2}) \log\left(1 + \frac{\sigma_{MF}^2}{\sigma_n^2}\right) \qquad (1)$$

$$I_{MF} = \sum_{\substack{Filter\\Channels}} \log\left(1 + \frac{\sigma_{MF}^2}{\sigma_n^2}\right) \qquad (2)$$

Where the σ_{MF}^2 is the variance in motion-free image after convolution with kernels designed to mimic the frequency response of the human visual system (HVS), σ_v^2 quantifies the variance introduced by motion artifacts, and σ_n^2 represents the noise in the HVS channel. The term g provides an estimate of the degradation in image information due to patient motion, which depends on the covariance between the motion-corrupted and motion-free images. All terms in Eq. 1 and Eq. 2 were set and calculated as described in Ref [3].

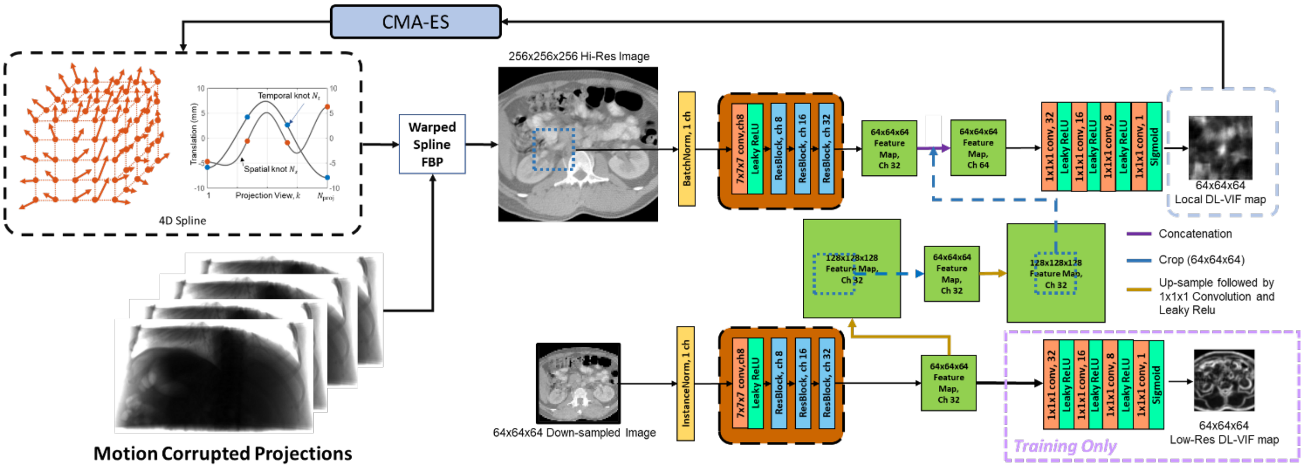

Figure 1. The network architecture of CADL-VIF Map and its integration into the deformable motion compensation framework.

The new VIF$_L$ was reproduced with a novel deep CNN based on our previous DL-VIF design and illustrated in Fig. 1. The network acts on small regions of interest (ROIs) of size 64x64x64 voxels (1 mm isotropic voxel size). Context information is incorporated via a second branch acting on the entire motion-corrupted volume reconstructed at a very coarse voxel size (4 mm isotropically), resulting in a multi-resolution, context aware architecture, inspired by previous work on CT to MR image synthesis [8]. The two branches featured identical layer configurations, with an input 7x7x7 convolution layer and a leaky-ReLU activation, followed by 3 ResBlock layers [3]. Thus, both branches contained an equal number of learnable parameters, yet independently learned. Both branches output 32-channel feature maps, one incorporating local, high-frequency features, and the other providing contextual feature information.

The high-resolution branch incorporated an input batch normalization layer, while the low-resolution branch featured instance normalization. The different normalization responds to different variability presented by the high and low-resolution data. The high-resolution ROIs featured variable soft-tissue and bone regions that present a much larger variability and benefit from batch normalization, compared to the relatively consistent appearance of the low resolution, full abdomen, context.

The context feature map from low-resolution branch was then up-sampled twice and cropped accordingly to match the position and size of the high-resolution ROI. The local and context feature maps were concatenated and input to cascade of 1x1x1 convolution and leaky-ReLU layers to generate the output of high-resolution branch. To facilitate contextual feature learning of the network during training, another series of 1x1x1 convolution with leaky ReLU layers were added to the low-resolution branch after the feature maps, generating low resolution output for the entire volume.

2.2 Deformable Motion Compensation Framework

CADL-VIF was incorporated as the autofocus metric in a multi-region-based motion compensation framework (see Fig. 1). The time-varying motion vector field (MVF) was estimated with a 4D spline model, integrated into the backprojector algorithm. Deformable motion was estimated as the set of 4D spline coefficients P minimizing the multi-ROI autofocus function:

$$P = \ argmin_P \sum_{r \in ROIs} \sum_{pixels} -\ln\left[S(\mu_{LR}(P), \mu_{HR}(P, r))\right] \qquad (3)$$

Where S is the autofocus metric for high-resolution ROI, i.e., the high-resolution CADL-VIF Map, calculated from $\mu_{LR}(P)$, the low-resolution image reconstructed with P, and $\mu_{HR}(P, r)$, the high-resolution image reconstructed with P at ROI position r. The negative natural logarithm served as a basic conditioning of the values for optimization. The final autofocus metric was integrated across ROI voxels and across all ROIs. The cost-function was minimized with the Covariance Matrix Adaptation Evolutionary Strategy (CMA-ES) [9].

2.3 Data Generation and Training

The CADL-VIF network was trained and validated on simulated data generated using 75 cases from the TCIA lymph node abdomen multi-detector CT (MDCT) database. 61 cases were used for training, 7 cases reserved for validation, and the rest 7 cases for testing. For each simulation instance, a MDCT volume was randomly selected and a 260 mm long sub-volume at a random longitudinal position was extracted from the original volume. The sub-volume was then forward projected using a high-fidelity CBCT projector with geometry pertinent to interventional robotic C-arm systems (source-to-detector distance of 1200 mm, and source-to-axis distance of 785mm). The detector was modeled as a flat-panel with 864 x 660 pixels and 0.64 mm isotropic pixel size. Deformable motion was induced during forward projection, using a MVF with random maximum amplitude ranging from 10 mm to 30 mm at random directions. The MVF featured maximum amplitude at a randomly placed position and decayed smoothly following an elliptical Gaussian kernel with lateral and antero-posterior width randomly chosen from 200 mm to 300 mm and 100 mm to 150 mm, respectively. The MVF followed a cosine temporal pattern with random phase and random frequency ranging from 0.75-1.25 cycles per scan. An additional motion-free scan was simulated to obtain a reference for ground truth $\mathrm{VIF_L}$. Both the motion-corrupted and motion-free images were reconstructed on a 256x256x256 voxels volume with 1x1x1 mm^3 voxel size.

Ground truth $\mathrm{VIF_L}$ maps of motion-corrupted images for training were computed using the motion-free images as reference. To emphasize motion artifacts in soft tissue regions, a [0.01, 0.025] mm^{-1} window, followed by CLAHE contrast-enhancement were applied to the motion-corrupted and motion-free volumes before calculation of $\mathrm{VIF_L}$. A total of 610 motion instances were generated for training, 70 for validation, and 70 for testing.

During training, the input data was normalized to [0,1], and data augmentation was achieved via addition of zero mean Gaussian noise with $\sigma = 0.01$. For each training instance pair, a randomly placed 64x64x64 voxels sub-volume was extracted from the motion corrupted image for input to the high-resolution branch, while the full volume, downsampled by a 4x factor, provided the low-resolution contextual input. Training was achieved with a loss function based on the mean square error between the network output and the ground truth $\mathrm{VIF_L}$. The network was trained with the Adam optimizer ($5\mathrm{x}10^{-4}$ learning rate), with a batch size of 30, and for 2000 epochs, on 3 Nvidia Quadro RTX A6000 GPUs.

2.4 Validation Experiments

To test the generalizability of CADL-VIF, a separate dataset with 256 cases was created from the test TCIA volumes, using a larger range of motion frequency: 0.5 to 3 periods per scan. CADL-VIF was estimated on contiguous ROIs of 64 x 64 x 64 voxels, covering the complete 256 mm x 256 mm x 256 mm volume. The sum of CADL-VIF within each ROI was then compared with the sum of ground truth $\mathrm{VIF_L}$ in the same region. Results were aggregated from all 64 ROIs in each of 256 cases.

Motion compensation with CADL-VIF was evaluated on 7 simulated cases created analogously to the test dataset, with motion amplitude ranging from 8 to 15 mm, and frequency ranging from 0.8 to 1.2 periods per scan. Motion compensation was performed with a total of 4 local ROIs of 64 mm x 64 mm x 64 mm size, with a common contextual reference, and using a 9x9x9x5 4-dimensional b-spline grid. Results were assessed with SSIM and blurriness estimated using the formula proposed in Ref [10], adapted to 3D, which is calculated as follows: i) each voxel in motion-free and motion-corrupted images is compared with all its neighboring 26 voxels, and the maximum intensity variation is stored in two new volumes, $\mathrm{V_{MF}}$ and $\mathrm{V_{MC}}$ for the motion-free and motion-corrupted images, respectively; ii) The average value of $\mathrm{V_{MF}}$ and $\mathrm{V_{MC}}$ is calculated (Z_{MF} and Z_{MC}, respectively); iii) blurriness is defined as $B = \frac{|Z_{MF}-Z_{MC}|}{Z_{MF}}$.

SIM was computed on 6mm x 6mm x 6mm ROIs, that underwent rigid registration to avoid phase mismatches between the motion-corrupted volume, compensated volume, and the static reference. Two ROIs were used per volume, one placed at the center of the motion field and a second one at a quasi-static region.

3. RESULTS

3.1 DL-VIF Map and Deformable Motion Severity

Figure 2 (A) shows the variation in $\mathrm{VIF_L}$ as a function of motion amplitude and motion frequency, to validate is capability to accurately quantify the effects of motion on image patches. $\mathrm{VIF_L}$ values show decreasing trend with increased amplitude (larger motion) and motion frequency (faster motion), making it a suitable metric to quantify motion induced image quality degradation. Validation of the capability of $\mathrm{VIF_L}$ to quantify motion is accompanied by assessment of the agreement between reference-based $\mathrm{VIF_L}$ maps and CADL-VIF inferences, illustrated in Fig. 2 (B). CADL-VIF showed good agreement with reference $\mathrm{VIF_L}$ across volumes throughout the entire extended dataset, achieving a slope of 0.941 and a linearity of 0.843. Combining these two results: i) $\mathrm{VIF_L}$ can accurately reflect the severity of local motion-induced image quality degradation for an extended range of deformable MVF; ii) the combination of contextual and local features learned with CADL-VIF are representative of motion artifacts decreasing the local value of $\mathrm{VIF_L}$, making CADL-VIF a suitable metric for reference-free motion quantification.

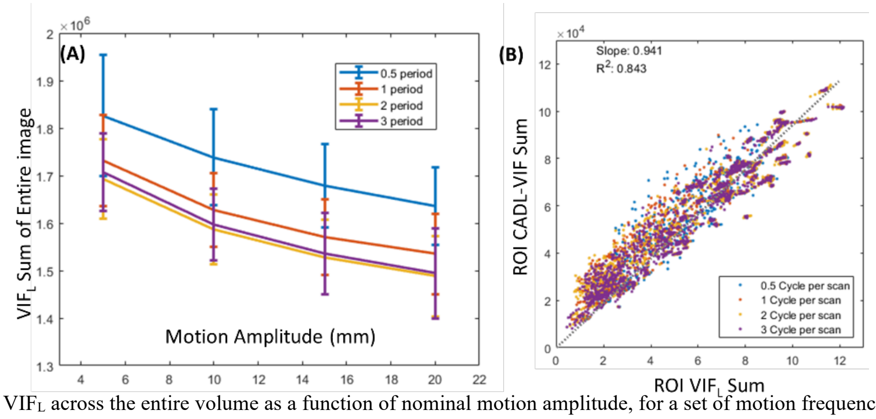

Figure 2. (A) Sum of $\mathrm{VIF_L}$ across the entire volume as a function of nominal motion amplitude, for a set of motion frequency values. (B) Agreement between the reference-free CADL-VIF inferences and conventional, reference-based VIFL, both integrated over 64x64x64-voxel

3.2 Motion Compensation with CADL-VIF Map

Figure 3 quantitatively illustrate the performance of deformable autofocus based on CADL-VIF for compensation of motion using a multi-ROI approach. CADL-VIF autofocus resulted net improvement in SSIM for all 7 motion cases, yielding an average increase in SSIM of 0.129 for regions with severe motion and 0.113 for regions with mild motion. It is worth noting that the seemingly low SSIM is likely due to the low contrast in soft tissue and the presence of noise, and it is the increase in SSIM value that reflects improved image quality, as illustrated in Figure 4, which shows an example compensation of severely motion-distorted anatomy. Figure 3 (B) show the reduction of image blurriness after motion compensation with CADL-VIF autofocus, with an average reduction of 0.032 in severe motion regions and 0.052 in mild motion regions, which is expected in successful motion compensation.

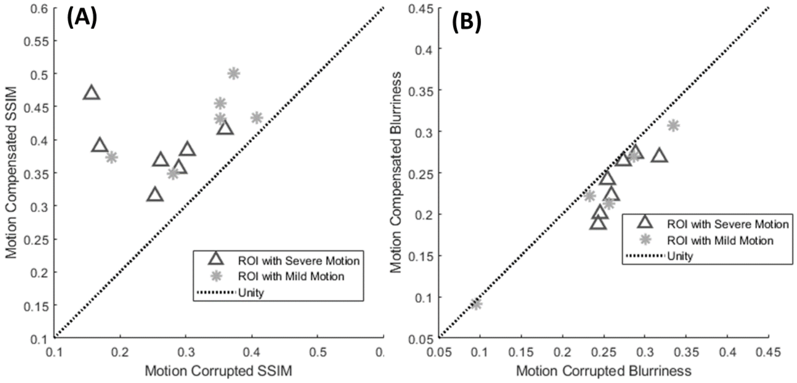

Figure 3. SSIM values (A) and blurriness (B) before and after motion compensation for regions with severe (black triangles) and mild motion (grey stars). The dashed lines mark the unity line, in which motion compensated and motion-corrupted images are equivalent. Values above the unity line in (A) indicate a net increase in SSIM after motion compensation. Values below the unity line in (B) are associated to sharper images.

Consistent improvement in SSIM and reduction in blurriness can be qualitatively appreciated in Figure 4. Motion artifacts severely distorted anatomical structures, with noticeable impact at the center of the volume (region of larger motion amplitude). The distortion and blurriness resulted in a sizable reduction of SSIM, to a value of 0.157. Autofocus motion compensation with CADL-VIF successfully restored the appearance of anatomical structures, mitigated shape distortion, and reduced image blurring, yielding a 3-fold increase in SSIM to 0.469.

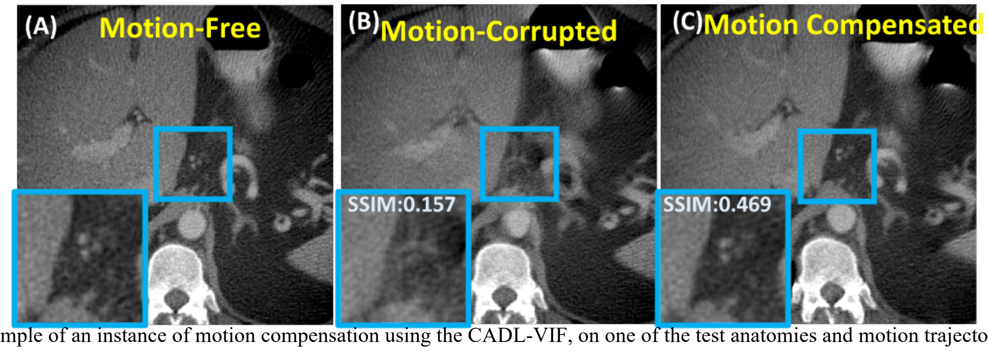

Figure 4. Example of an instance of motion compensation using the CADL-VIF, on one of the test anatomies and motion trajectories.

4. DISCUSSION AND CONCLUSION

This work presents a new learning-based image quality metric (CADL-VIF) to quantify the effect of CBCT deformable motion within a local region of interest. The proposed network builds on our previous approach for rigid autofocus with learned metrics providing simultaneous quantification of image quality and structural integrity of the underlying anatomy, by integrating contextual information and extending the reference similarity metrics from scalar values to spatially varying distributions for generation of voxel-wise maps of motion artifacts and distortion on high-resolution ROIs. Such locality is crucial for integration into multi-ROI deformable autofocus compensation methods that would otherwise be unfeasible due to large computational and memory requirements of performing the compensation in the complete volume, while the integration of contextual information allows robust estimation of motion effects by mitigating the effect of confusing factors associated to local variations of image content.

CADL-VIF was able to accurately reproduce the motion quantification capability of the reference similarity metric. When integrated into the CBCT deformable autofocus framework, CADL-VIF proved capable of recovering fine details in soft tissue structures that challenge conventional metrics. Ongoing work targets application of CADL-VIF to clinical data scenarios via training with simulated datasets including complete models of the CBCT imaging chain.

ACKNOWLEGEMENT

This work was supported by the National Institute of Health, Grant R01-EB-030547.

REFERENCES

[1] Sisniega A, Stayman J W, Yorkston J, Siewerdsen J H and Zbijewski W 2017 Motion compensation in extremity cone-beam CT using a penalized image sharpness criterion *PMB* **62** 3712–34, https://doi.org/10.1088/1361-6560/aa6869

[2] Capostagno S, Sisniega A, Stayman J W, Ehtiati T, Weiss C R and Siewerdsen J H 2021 Deformable motion compensation for interventional cone-beam CT *Phys. Med. Biol.* **66** 055010, https://doi.org/10.1088/1361-6560/abb16e

[3] Huang H, Siewerdsen J H, Zbijewski W, Weiss C R, Unberath M, Ehtiati T and Sisniega A 2022, Reference-free learning-based similarity metric for motion compensation in cone-beam CT, *Phys. Med. Biol.* **67** 125020, https://doi.org/10.1088/1361-6560/ac749a

[4] Huang H, Siewerdsen J H, Zbijewski W, Weiss C R, Ehtiati T and Sisniega A 2021 Reference-free, learning-based image similarity: application to motion compensation in cone-beam CT *16th Virtual Int. Meeting on Fully 3D Image Reconstruction in Radiology and Nuclear Medicine* pp 67–71, http://arxiv.org/abs/2110.04143

[5] Sheikh H R and Bovik A C 2006 Image information and visual quality *TIP* **15** 430–44, https://doi.org/10.1109/TIP.2005.859378

[6] Sisniega A, Huang H, Zbijewski W, Stayman J W, Weiss C R, Ehtiati T and Siewerdsen J H 2021 Deformable image-based motion compensation for interventional cone-beam CT with a learned autofocus metric *Proc SPIE* **11595** 115950W, https://doi.org/10.1117/12.2582140

[7] Shao, Y, Sun, F, Li, H and Liu, Y 2014. A Novel Approach for Computing Quality Map of Visual Information Fidelity Index. In: Sun, F., Li, T., Li, H. (eds) Foundations and Applications of Intelligent Systems. Advances in Intelligent Systems and Computing, vol 213. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37829-4_14

[8] Kläser K, Borges P, Shaw R, Ranzini M, Modat M, Atkinson D, Thielemans K, Hutton B, Goh V, Cook G, Cardoso J and Ourselin S 2021 A multi-channel uncertainty-aware multi-resolution network for MR to CT synthesis *Appl. Sci.* **11** 1667, https://doi.org/10.3390/app11041667

[9] Hansen N 2014 CMA-ES: a function value free second order optimization method, https://hal.inria.fr/hal-01110313

[10] Elsayed M, Sammani F, Hamdi A, Albser A, Babalghoom H 2018 A New Method for Full Reference Image Blur Measure, Int J Simul Sci Technol. 19(1):4, http://dx.doi.org/10.5013/IJSSST.a.19.01.7