Enabling Geometry Aware Learning Through Differentiable Epipolar View Translation

Maximilian Rohleder[1,2(✉)], Charlotte Pradel[1], Fabian Wagner[1], Mareike Thies[1], Noah Maul[1,2], Felix Denzinger[1,2], Andreas Maier[1], and Bjoern Kreher[2]

[1] Friedrich-Alexander-University, Erlangen-Nürnberg, Erlangen, Germany
Maxi.Rohleder@fau.de
[2] Siemens Healthineers AG, Erlangen, Germany

Abstract. Epipolar geometry is exploited in several applications in the field of Cone-Beam Computed Tomography (CBCT) imaging. By leveraging consistency conditions between multiple views of the same scene, motion artifacts can be minimized, the effects of beam hardening can be reduced, and segmentation masks can be refined. In this work, we explore the idea of enabling deep learning models to access the known geometrical relations between views. This implicit 3D information can potentially enhance various projection domain algorithms such as segmentation, detection, or inpainting. We introduce a differentiable feature translation operator, which uses available projection matrices to calculate and integrate over the epipolar line in a second view. As an example application, we evaluate the effects of the operator on the task of projection domain metal segmentation. By re-sampling a stack of projections into orthogonal view pairs, we segment each projection image jointly with a second view acquired roughly 90° apart. The comparison with an equivalent single-view segmentation model reveals an improved segmentation performance of 0.95 over 0.91 measured by the dice coefficient. By providing an implementation of this operator as an open-access differentiable layer, we seek to enable future research.

Keywords: Cone-Beam Computed Tomography · Epipolar Geometry · Operator Learning

1 Introduction

Cone-Beam Computed Tomography (CBCT) scans are widely used for guidance and verification in the operating room. The clinical value of this imaging modality is however limited by artifacts originating from patient and device motion or metal objects in the X-Ray beam. To compensate these effects, knowledge about the relative acquisition geometry between views can be exploited. This so-called epipolar geometry is widely used in computer vision and can be applied to CBCT imaging due to the similar system geometry [3,11].

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14222, pp. 57–65, 2023.
https://doi.org/10.1007/978-3-031-43898-1_6

By formulating and enforcing consistency conditions based on this geometrical relationship, motion can be compensated [4,8], beam-hardening effects can be reduced [11], and multi-view segmentations can be refined [2,7]. Motion and beam hardening effects can be corrected by optimizing for consistency through either updating the projection matrices or image values while the respective other is assumed fixed.

In segmentation refinement, the principal idea is to incorporate the known acquisition geometry to unify the binary predictions on corresponding detector pixels. This inter-view consistency can be iteratively optimized to reduce false-positives in angiography data [8]. Alternatively, an entire stack of segmented projection images can be backprojected, the reconstructed volume thresholded and re-projected to obtain 3D consistent masks [2].

In this work, we explore the idea of incorporating epipolar geometry into the learning-based segmentation process itself instead of a separate post-processing step. A differentiable image transform operator is embedded into the model architecture, which translates intermediate features across views allowing the model to adjust its predictions to this conditional information. By making this information accessible to neural networks and enabling dual-view joint processing, we expect benefits for projection domain processing tasks such as inpainting, segmentation or regression. As a proof-of-concept, we embed the operator into a segmentation model and evaluate its influence in a simulation study. To summarize, we make the following contributions:

- We analytically derive formulations for forward- and backward pass of the view translation operator
- We provide an open-source implementation thereof which is compatible with real-world projection matrices and PyTorch framework
- As an example of its application, we evaluate the operator in a simulation study to investigate its effect on projection domain segmentation

2 Methods

In the following sections we introduce the geometrical relationships between epipolar views, define a view translation operator, and analytically derive gradients needed for supervised learning.

Epipolar Geometry and the Fundamental Matrix. A projection matrix $P \in \mathbb{R}^{3\times4}$ encodes the extrinsic device pose and intrinsic viewing parameters of the cone-beam imaging system. These projection matrices are typically available for images acquired with CBCT-capable C-Arm systems. Mathematically, this non-linear projective transform maps a point in volume coordinates to detector coordinates in homogeneous form [1]. When two projection images of the same scene are available, the two detector coordinate systems can be linked through epipolar geometry as depicted in Fig. 1. The Fundamental matrix $F \in \mathbb{R}^{3\times3}$ directly encodes the inherent geometric relation between two detector coordinate

systems. More specifically, a point $\mathbf{u}'$ in one projection image is mapped onto a line $\mathbf{l}$ through $\mathbf{l} = F\mathbf{u}'$, where $\mathbf{l}, \mathbf{u}$ are vectors in the 2D projective homogeneous coordinate space $\mathbb{P}^{2+}$ (notation from [1]). Given projection matrices $P, P' \in \mathbb{R}^{3\times4}$, the fundamental matrix can be derived as

$$F = [P\mathbf{c}']_\times P P'^{+} \quad , \tag{1}$$

where $\cdot^+$ denotes the pseudo-inverse, $\mathbf{c}' \in \mathbb{P}^{3+}$ is the camera center in homogeneous world coordinates, and $[\cdot]_\times$ constructs the tensor-representation of a cross product. Note, that the camera center can be derived as the kernel of the projection $\mathbf{c} = \ker(P)$. Additional details on epipolar geometry can be found in literature [3].

The Epipolar View Translation Operator (EVT). The goal of the proposed operator is to provide a neural network with spatially registered feature information from a second view of known geometry. Consider the dual view setup as shown in Fig. 1. Epipolar geometry dictates that a 3D landmark detected at a detector position $\mathbf{u}'$ in projective view P' is located somewhere along the epipolar line $\mathbf{l}$ in the respective other view P. Naturally this only holds true as long as the landmark is within the volume of interest (VOI) depicted in both images.

To capture this geometric relationship and make spatially corresponding information available to the model, an epipolar map Ψ' is computed from the input image p. As shown in Eq. 2, each point $\mathbf{u}'$ in the output map Ψ', is computed as the integral along its epipolar line $\mathbf{l} = F\mathbf{u}'$.

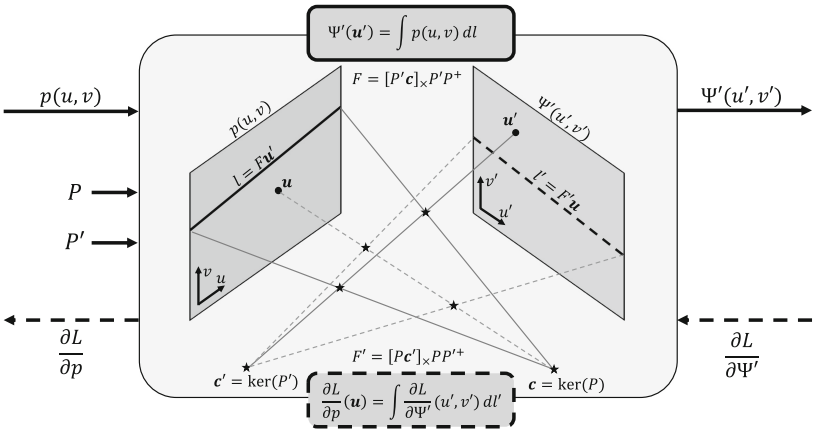

Fig. 1. Epipolar Geometry as defined in [1,3] and the proposed View Translation Operator. The intrinsic relation between two images under projection P and P' is compactly captured in the Fundamental matrix F. In the forward pass, a point $\mathbf{u}'$ in the epipolar map Ψ is calculated by integrating along the epipolar line $\mathbf{l} = F\mathbf{u}'$. During gradient backpropagation, the contribution of a point $\mathbf{u}$ in the input image is determined by integrating along its epipolar line in the gradient image $\frac{\partial L}{\partial \Psi'}$.

$$\Psi'(\mathbf{u}') = \int_L p(\mathbf{u})\, dl, \quad L := \{\mathbf{u} \in \mathbb{P}^{2+} : \mathbf{u}^\top F \mathbf{u}' = 0\} \tag{2}$$

Gradient Derivation. To embed an operator into a model architecture, the gradient with respect to its inputs and all trainable parameters needs to be computed. As the proposed operator contains no trainable parameters, only the gradient with respect to the input is derived.

The forward function for one output pixel $Y_{\mathbf{u}'} = \Psi'(\mathbf{u}')$ can be described through a 2D integral over the image coordinates $\mathbf{u}$

$$Y_{\mathbf{u}'} = \iint_{\mathbf{u}} \delta(\mathbf{u}', \mathbf{u}, F) X_{\mathbf{u}} \ , \tag{3}$$

where $X_{\mathbf{u}}$ denotes the value in the input image $p(u, v)$ at position $\mathbf{u}$. Here, the indicator function $\delta(\cdot)$ signals if the coordinate $\mathbf{u}$ lies on the epipolar line defined by F and $\mathbf{u}'$:

$$\delta(\mathbf{u}', \mathbf{u}, F) = \begin{cases} 1, & \mathbf{u}^\top F \mathbf{u}' = 0 \\ 0, & \text{else} \end{cases} \tag{4}$$

After calculation of a loss L, it is backpropagated through the network graph. At the operator, the loss arrives w.r.t. to the predicted consistency map $\frac{\partial L}{\partial \mathbf{Y}}$. From this image-shaped loss, the gradient w.r.t. the input image needs to be derived. By marginalisation over the loss image $\frac{\partial L}{\partial \mathbf{Y}}$, the contribution of one intensity value $X_{\mathbf{u}}$ in the input image can be written as

$$\frac{\partial L}{\partial X_{\mathbf{u}}} = \frac{\partial L}{\partial \mathbf{Y}} \frac{\partial \mathbf{Y}}{\partial X_{\mathbf{u}}} = \iint_{\mathbf{u}'} \frac{\partial L}{\partial Y_{\mathbf{u}'}} \frac{\partial Y_{\mathbf{u}'}}{\partial X_{\mathbf{u}}} \ . \tag{5}$$

Deriving Eq. 3 w.r.t. $X_{\mathbf{u}}$ eliminates the integral and only leaves the indicator function as an implicit form of the epipolar line. With this inserted in Eq. 5, the loss for one pixel in the input image can be expressed as

$$\frac{\partial L}{\partial X_{\mathbf{u}}} = \iint_{\mathbf{u}'} \delta(\mathbf{u}', \mathbf{u}, F) \frac{\partial L}{\partial Y_{\mathbf{u}'}} \ . \tag{6}$$

Note, that forward (Eq. 3) and backward formulation (Eq. 6) are similar and can thus be realised by the same operator. Due to the symmetry shown in Fig. 1, the backward function can be efficiently formulated similar to Eq. 3 by integration along the line $\mathbf{l}'$ defined by the reversed Fundamental matrix F' as

$$\frac{\partial L}{\partial p}(\mathbf{u}) = \int_{L'} \frac{\partial L}{\partial \Psi'}(\mathbf{u}')\, dl' \quad L' := \{\mathbf{u}' \in \mathbb{P}^{2+} : \mathbf{u}'^\top F' \mathbf{u} = 0\} \ . \tag{7}$$

Implementation. The formulations above used to derive the gradient assume a continuous input and output distribution. To implement this operation on discrete images, the integral is replaced with a summation of constant step size of one pixel and bi-linear value interpolation. As the forward and backward functions compute the epipolar line given the current pixel position, slight mismatches in interpolation coefficients might occur. However, well-tested trainable reconstruction operators use similar approximate gradients and are proven to converge regardless [6]. The differentiable operator is implemented as a *PyTorch* function using the `torch.utils.cpp_extension`. To enable parallel computation, the view transformation is implemented as a CUDA kernel. The source code will be made public upon publication.[1]

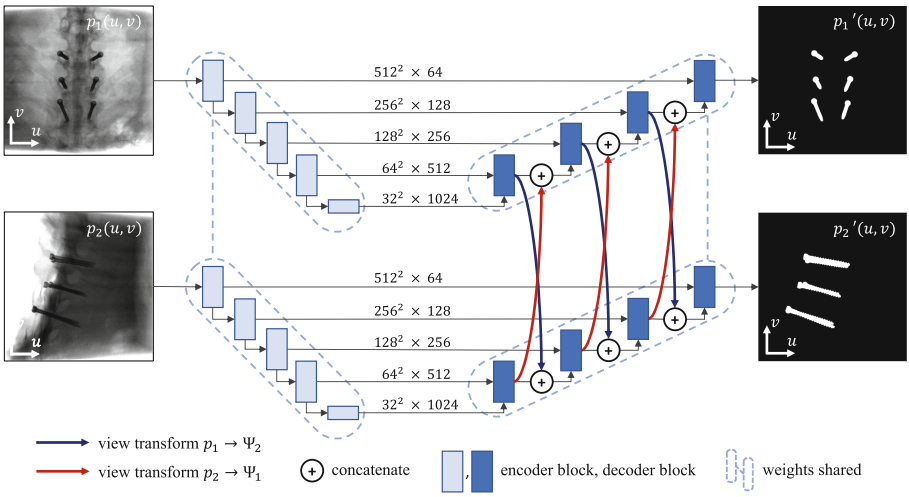

Fig. 2. Dual View Segmentation Model with embedded EVT operator. The operator is embedded at three different scale-levels in the decoder section of a 2D U-Net [5]. By exchanging spatially translated feature maps after each up-block features from the second view are considered during mask decoding.

3 Experiments

As a proof-of-concept application, we assess the operator's influence on the task of projection domain metal segmentation. To reconstruct a CBCT scan, 400 projection images are acquired in a circular trajectory over an angular range of 200°. To experimentally validate the effects of the proposed image operator, the images are re-sampled into orthogonal view pairs and jointly segmented using a model with an embedded EVT operator.

[1] https://github.com/maxrohleder/FUME.

Dual View U-Net with EVT Operator. To jointly segment two projection images, the Siamese architecture shown in Fig. 2 is used. It comprises two U-Net backbones with shared weights which process the two given views. The EVT operator is embedded as a skip connection between the two mirrored models to spatially align feature maps with the respective other view. Each view is fed through the model individually up to the point where epipolar information is added. There, the forward pass is synchronized and the translated feature maps from the respective other view are concatenated. We place the operator at three positions in the model – right after each upsampling block except the last. In the decoder block, feature maps are arguably more sparse. Intuitively, this increases the value of the proposed operator as fewer objects are in the same epipolar plane and thus correspondence is more directly inferable.

Compared Models. To investigate the effects of the newly introduced operator, the architecture described above is compared to variants of the U-Net architecture. As a logical baseline, the plain U-Net architecture from [5] is trained to segment each projection image individually. Additionally, the same architecture is provided with two projection images concatenated along the channel axis. This approach verifies that any changes in segmentation performance are attributable to the feature translation operator and not simply due to providing a second view.

Data. For the purpose of this study, we use a simulated projection image dataset. Analogous to DeepDRR [9], we use an analytical polychromatic forward model to generate X-Ray images from given CBCT volume data. In total, 29 volumes with approximate spatial dimensions $16\,\mathrm{cm}^3$ from 4 anatomical regions are used ($18\times$ spine, $4\times$ elbow, $5\times$ knee, and $2\times$ wrist). To simulate realistic metal shapes, objects are selected from a library of surgical tools made available by Nuvasive (San Diego, USA) and Königsee Implantate (Allendorf, Germany).

From this primary data, six spine scans are selected for testing, and three are selected for validation. Metal implants are manually positioned relative to the anatomy using 3D modelling tools. The metal objects are assembled such that they resemble frequently conducted procedures including pedicle screw placement and k-wire insertions. In total, there are 12 unique scenes fitted to the scans in the test set, and 5 in the validation set.

The training set consists of randomly selected and assembled metal objects. Each of the remaining 20 volumes is equipped with $n \in \{4, 6, 8, 10\}$ randomly positioned (non-overlapping) metal objects creating 80 unique scenes.

During simulation, the objects are randomly assigned either iron or titanium as a material which influences the choice of attenuation coefficients. For each scene, 100 projection images are generated whose central ray angles on the circular trajectory are approximately 2° apart.

3.1 Model Training

The three models are trained using the Adam optimizer with the dice coefficient as a loss function and a learning rate of 10^{-5}. The best model is selected on the validation loss. As a data augmentation strategy, realistic noise of random strength is added to simulate varying dose levels or patient thickness [10]. We empirically choose the range of noise through the parameter photon count $\#p \in [10^2, 10^4]$. During validation and testing, the noise is set to a medium noise level $\#p = 10^3$. Furthermore, each projection image is normalized to its own mean and standard deviation. The models are trained for 200 epochs on an NVIDIA A100 graphics card.

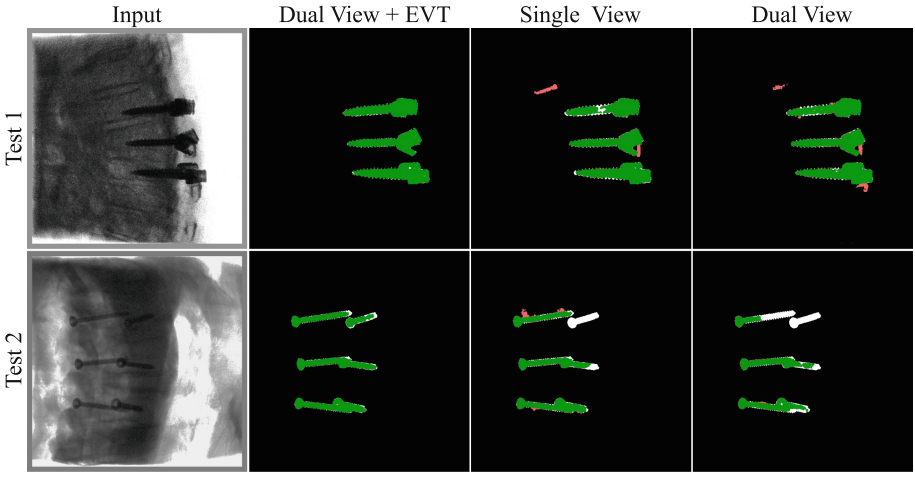

Fig. 3. Segmentation results from the three compared models on three selected projection images of the test set. Correctly segmented pixels are colored green and black, white indicates false negatives and red indicates the false positives. (Color figure online)

4 Results

Quantitative. The per-view averaged segmentation test set statistics are shown in Table 1. The model predicting a single view yields a dice score of 0.916 ± 0.119, thus outperforming the model which is fed two projection images at an average dice similarity of 0.889 ± 0.122. The model equipped with our operator, which also is presented with two views, but translates feature maps internally, yields the highest dice score of 0.950 ± 0.067.

Qualitative. To illustrate the reported quantitative results, the segmentation prediction is compared on two selected projection images in Fig. 3. Test 1 shows a lateral view of a spine with 6 pedicle screws and tulips inserted into the 3 central

Table 1. Quantitative evaluation of the three compared model variations over the test set. All values reported as mean ± standard deviation.

	Dice	Precision	Recall
Single View	0.916 ± 0.119	0.976 ± 0.013	0.882 ± 0.168
Dual View	0.889 ± 0.122	0.970 ± 0.013	0.842 ± 0.177
Dual View + EVT	**0.950 ± 0.067**	**0.991 ± 0.010**	**0.919 ± 0.106**

vertebrae. It is shown here as an approximately representative illustration of the improvement in precision due to the operator. A bone edge is falsely segmented as metal in both the single view and dual view mode. The model equipped with the operator neglects this false segmentation. Test 2 illustrates a case of improved recall, where one screw (top right) is occluded by strongly attenuating anatomy. Our model is able to recover the screw at least partially, whereas it is missed by the two other models.

5 Discussion and Conclusion

Building upon previous work on epipolar geometry and the resulting consistency conditions in X-Ray imaging, we propose a novel approach to incorporate this information into a model. Rather than explicitly formulating the conditions and optimizing for them, we propose a feature translation operator that allows the model to capture these geometric relationships implicitly.

As a proof-of-concept study, we evaluate the operator on the task of projection domain segmentation. The operator's introduction enhances segmentation performance compared to the two baseline methods, as shown by both qualitative and quantitative results. Primarily we found the information from a second view made available by the operator to improve the segmentation in two ways: (1) Reduction of false positive segmentations (2) Increased sensitivity in strongly attenuated areas. Especially for the segmentation of spinal implants, the model performance on lateral images was improved by epipolar information from an orthogonal anterior-posterior projection. Lateral images are usually harder to segment because of the drastic attenuation gradient as illustrated in Fig. 3. It is noteworthy that the U-Net architecture utilizing two images as input exhibits inferior performance compared to the single view model. The simple strategy of incorporating supplementary views into the network fails to demonstrate any discernible synergistic effect, likely due to the use of the same model complexity for essentially conducting two segmentation tasks simultaneously.

The simulation study's promising results encourage exploring the epipolar feature transform as a differentiable operator. Future work involves evaluating the presented segmentation application on measured data, analyzing the optimal integration of the operator into a network architecture, and investigating susceptibility to slight geometry calibration inaccuracies. As the U-Net's generalization on real data has been demonstrated, we expect no issues with our method.

In conclusion, this work introduces an open-source[2] differentiable operator to translate feature maps along known projection geometry. In addition to analytic derivation of gradients, we demonstrate that these geometry informed epipolar feature maps can be integrated into a model architecture to jointly segment two projection images of the same scene.

Data Use Declaration. The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects, or cadavers of both kind. All data acquisitions were done in consultation with the Institutional Review Board of the University Hospital of Erlangen, Germany.

References

1. Aichert, A., et al.: Epipolar consistency in transmission imaging. IEEE Trans. Med. Imaging **34**(11), 2205–2219 (2015). https://doi.org/10.1109/TMI.2015.2426417
2. Gottschalk, T.M., Maier, A., Kordon, F., Kreher, B.W.: Learning-based patch-wise metal segmentation with consistency check. In: Bildverarbeitung für die Medizin 2021. I, pp. 4–9. Springer, Wiesbaden (2021). https://doi.org/10.1007/978-3-658-33198-6_4
3. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2004). https://doi.org/10.1017/CBO9780511811685
4. Preuhs, A., et al.: Symmetry prior for epipolar consistency. IJCARS **14**(9), 1541–1551 (2019). https://doi.org/10.1007/s11548-019-02027-8
5. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
6. Syben, C., Michen, M., Stimpel, B., Seitz, S., Ploner, S., Maier, A.K.: Technical Note: PYRONN: Python reconstruction operators in neural networks. Med. Phys. **46**(11), 5110–5115 (2019). https://doi.org/10.1002/mp.13753
7. Unberath, M., Aichert, A., Achenbach, S., Maier, A.: Improving segmentation quality in rotational angiography using epipolar consistency. In: Balocco, S. (ed.) Proc MICCAI CVII-STENT, Athens, pp. 1–8 (2016)
8. Unberath, M., Aichert, A., Achenbach, S., Maier, A.: Consistency-based respiratory motion estimation in rotational angiography. Med. Phys. **44**(9), e113–e124 (2017)
9. Unberath, M., et al.: DeepDRR – a catalyst for machine learning in fluoroscopy-guided procedures. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11073, pp. 98–106. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_12
10. Wang, A., et al.: Low-dose preview for patient-specific, task-specific technique selection in cone-beam CT. Med. Phys. **41**(7), 071915 (2014). https://doi.org/10.1118/1.4884039
11. Würfl, T., Hoffmann, M., Aichert, A., Maier, A.K., Maaß, N., Dennerlein, F.: Calibration-free beam hardening reduction in x-ray CBCT using the epipolar consistency condition and physical constraints. Med. Phys. **46**(12), e810–e822 (2019). https://doi.org/10.1002/mp.13625

[2] https://github.com/maxrohleder/FUME.