

## Executive summary

Tomographic imaging techniques, like X-ray computed tomography (CT) and positron emission tomography (PET), aim to computationally recover the internal structure of an object from indirect observations that are obtained by probing the object with a wave/particle from multiple known directions. The *overall aim* of this project is to develop theory/algorithms for image reconstruction applicable to *spatiotemporal* and *multimodal/multichannel tomographic imaging*. This means addressing challenges arising when the intrinsic dynamic of the imaged object needs to be recovered during reconstruction (*spatiotemporal imaging*) and when the object is probed simultaneously with different waves/particles (*multimodal/multichannel imaging*). Previous efforts in this field deal only with post-processing, i.e., they do not address the intrinsic ill-posedness of (multichannel) image reconstruction. Our approach will be based on coupling channels and modelling dynamics with physics informed deep generative models. The methods will be generic, yet adaptable to specific use cases by incorporating the physics models through a plug-and-play scheme. This novel approach combines the power of generative models with the reliability of models derived from first principles. The work will be spearheaded by the two use cases: dynamic PET/CT, which requires incorporating the dynamics of both activity and anatomy, and dynamic photon-counting X-ray CT, which requires simultaneous handling of multichannel energy-resolved data and motion.

The foundation models for reconstruction in multichannel imaging will be based on a structure for incorporating handcrafted physics informed models that couple the channels to each other. This will allow unifying foundation models for spectral CT and dynamic PET/CT. PI 1 will lead the development of foundational models for image reconstruction in multichannel imaging that provide high-quality reconstructions by efficiently integrating complementary information from different channels. Small-scale multichannel imaging problems will be addressed at first and then focus will switch to scaling up the methods so that they can be applied to dynamic PET/CT and spectral CT. A starting point is to study transformer architectures for multichannel imaging based on cross attention [88] and architectures based on state space models (SSMs), like Mamba [33], that predict an output sequence from a continuous input sequence through learned state and output equations. PI 3 will extend the suitable deep neural network (DNN) architectures introduced by PI 1 to the dynamic PET setting and then adapt and apply techniques for pre-training and fine tuning against domain specific datasets with self-supervision and model balancing. The effect of substituting simple convolution in temporal domain with physics informed DNNs will also be explored, with the expectation of achieving models that generalise well from small training sets. Building on work by PIs 1 and 3, realistic simulations of line-of-response data and of biomarker kinematics will be generated and included in synthetic data sets. The methods will be tested on at least 200 clinical cases.

To apply the methods to spectral CT, PI 2 will establish a data platform and a digital twin for spectral CT by generating a library of physiologically realistic multichannel phantoms and use these to simulate dynamic, multispectral CT images. PI 2 will then extend the DNN architectures by PI 1 to a multichannel setting by introducing cross attention or cross-layer token fusion and include physics models to accommodate dynamic effects such as rigid motion, tissue compression, and blood flow and perfusion. We will evaluate the resulting algorithms on simulated data and on clinical data from a prototype photon-counting spectral CT scanner.

The application of deep learning to two use cases in medical imaging fits well with the research theme ‘Learn’ and the goal ‘Rich and Healthy Life’. With the proposed AI-tools we will enable dynamic PET as a routine clinical modality with a similar patient throughput of static PET and drastically improve CT imaging of cancer, stroke, and cardiovascular diseases, in addition to furthering the research frontier on DNN based approaches for multichannel imaging.

## State-of-the-art and scientific novelty

The *overall aim* is to develop theory/algorithms for image reconstruction applicable to *spatiotemporal* and *multimodal/multichannel imaging*. One challenge arises when the object being imaged has an intrinsic dynamics that needs to be recovered during reconstruction (*spatiotemporal imaging*). Another arises when the object is probed simultaneously with different waves/particles, so data or images have multiple channels (*multimodal/multichannel imaging*).

Much effort has been devoted to multimodal artificial intelligence (AI) in medical imaging [35, 45, 47, 1, 18, 24, 72] and there are some attempts at establishing foundational models as in [13] for 3D computed tomography (CT). However, *all of these deal with post-processing, i.e., they do not address the intrinsic ill-posedness that arises in (multichannel) image reconstruction*. The approach we take will be based on coupling channels and modelling dynamics with physics informed deep generative models. The methods will be generic, yet adaptable to specific use cases by incorporating the appropriate physics model through a plug-and-play scheme. This constitutes a novel approach for inverse problems that combines the power of generative models with reliability of models derived from first principles. The work will be spearheaded by the following two use cases.

**Dynamic PET/CT:** Positron emission tomography (PET) is a molecular imaging modality in which the coincident detection of photons from positron annihilations is used to image the bio-distribution of a  $\beta^+$ -labelled molecule [65]. Often, a CT scan is also performed immediately prior to the PET scan, in order to recover the 3D interior structure (anatomy map), which is needed to accurately model the PET data generated by an activity map (PET forward operator). Moreover, a typical PET data acquisition takes about 20 minutes during which some organs will move. The activity and anatomy maps are therefore time dependent. In modern scanners, coincidence sorting is performed *a posteriori*, and the recorded raw data is a list of photons with their energy and a time stamp. So-called list-mode data, i.e., coincidence events with time stamps, arrival time differences and energies, is then derived and used for image reconstruction. This allows investigating not only the bio-distribution of the tracer (activity map) but also the kinetics of the injected radiopharmaceutical [15] (dynamic PET). It also enables parametric PET imaging [64] that uses PET list-mode-data to reconstruct a steady-state activity map and other time-dependent, voxel-specific physiological parameters (e.g., diffusion from plasma to cells).

Dynamic PET/CT is thus a spatiotemporal multichannel imaging modality and accurate reconstruction of the activity channel requires recovering and incorporating the dynamics of the activity/anatomy from list-mode PET data. This is, for example, needed to increase spatial and contrast resolution in thoracic gated PET/CT imaging where the clinical goal is detecting lung tumours  $< 1\text{--}3$  cm in size [16] and for studying the kinetics of the radiopharmaceutical distribution, with potentially more valuable diagnostic power than standard PET [83] and wide-ranging applicability in drug and treatment development [64]. Notably, confounding organ movement and radiotracer kinematics can be attacked with similar strategies, which we aim to investigate.

Recently, much effort has been devoted to using deep learning for PET [68, 9, 58, 12, 37], and PET/CT reconstruction. [28, 23]. Principal investigator (PI) 3 has led development of AI models for PET/CT image reconstruction with promising results [34, 3, 2, 49, 48]. These are based on physics informed deep neural network (DNN) architecture introduced by PI 1 [5]. PIs 1 and 3 have also developed methods for spatiotemporal image reconstruction in gated PET/CT. PI 1 led work on methods that blend reconstruction with non-rigid registration with state-of-the-art (SOTA) results for low-dose spatiotemporal CT [20, 32, 19, 51, 40], which have been successfully applied by PIs 1 and 3 to PET/CT. To our knowledge, there are no documented attempts to use deep learning in projection data space for dynamic PET imaging [66, 34, 16].

**Spectral CT:** X-ray computed tomography (CT) is one of the most common medical imaging modalities. It reconstructs a 3D map of the patient from x-ray transmission measurements from multiple directions. Conventional CT scanners measure a single energy channel, yielding a ‘greyscale’ image, but cannot measure atomic composition and do not provide fully quantitative measurements due to beam-hardening artifacts that depend on scanner settings or surrounding anatomy. Dual-energy CT scanners can partially address these problems but suffer from imperfect iodine-calcium separation, and their clinical adoption has been slow.

Energy-resolving photon-counting detectors, recently introduced in the clinic, promise to overcome these limitations [22, 71, 30, 82, 81, 80, 67, 55] while also providing improved spatial resolution and better noise properties. Their energy-resolving measurements can be used for material decomposition, generating maps of material densities that can then be turned into quantitative images free of beam hardening. PI 2 has played an important role in developing a photon-counting prototype CT scanner based on a silicon detector, with a unique combination of high-count-rate performance, spatial resolution and energy resolution [63, 77, 62, 7, 70, 42].

Taking full advantage of photon-counting CT detectors will require reconstruction algorithms that can handle not only multichannel energy-resolved data, but also motion. If not corrected for, motion artifacts due to organ and blood vessel motion can degrade the quantitative accuracy of photon-counting CT images and degrade spatial resolution [31]. Conventional, analytical image-reconstruction methods are not able to handle noise and motion optimally.

PI 1 has pioneered work on several AI models for CT reconstruction as surveyed in [10, 60]. Many of these involve physics informed DNN architectures [5, 39, 59, 69] that have demonstrated SOTA performance [52, 69] with computationally feasible reconstruction time compared to variational models. PIs 1 and 2 have also worked on applying these and other AI models to spectral CT [26, 25, 78, 41, 44]. Results are encouraging, but several challenges remain to be solved before AI models for multichannel, spatiotemporal reconstruction in photon-counting spectral CT are ready for being implemented in clinical practice.

## Project plan

We will base the foundation models for reconstruction in multichannel imaging on a plug-and-play structure for incorporating physics informed models that couple the channels to each other. Since one can view spatiotemporal imaging as a special case of multichannel imaging by representing the temporal dimension as an additional channel, part of the work on foundation models for spectral CT and dynamic PET/CT can be unified. The project consists of three parts: developing general foundational models and applying these models to PET/CT and spectral CT.

**Foundational models:** The overall aim is to develop foundational models for image reconstruction in multichannel imaging that provide high-quality reconstructions by efficiently integrating complementary information from different channels. The work will be led by PI 1 and the first phase will consider small-scale multichannel imaging problems. The next phase focuses on scaling up the methods so that they can be applied to dynamic PET/CT and spectral CT. This part will be pursued in close collaboration with PIs 2 and 3.

A key part concerns developing physics informed DNN architectures for multichannel imaging. We will start by considering transformers that use attention to encode long-range dependencies. By varying the attention mechanism [38], one can adapt transformers to different tasks like natural language processing (NLP) [90], vision [36], image denoising [27] and reconstruction [2, 3, 49], the latter as part of unrolled architectures in work led by PI 3. The first step will seek to develop a transformer architecture for multichannel imaging based on cross

attention [88]. This and other model fusion techniques [89] have been successfully used in applying transformers for multimodal tasks involving NLP and vision [29, 92, 84, 54, 21, 91, 14]. A major drawback with transformers is that they have quadratic computational complexity. Various variants seek to address this [36], but setting up and training transformers can still become prohibitive in large-scale multichannel imaging problems, like clinical spectral CT. We will therefore consider state space models (SSMs), like Mamba [33], that predict an output sequence from a continuous input sequence through learned state and output equations. These DNN architectures can capture long-range dependencies with linear computational complexity [61, 6]. Recently they have been used in vision [86] and image reconstruction [43]. As with transformers, it is possible to use cross-layer token fusion [75] to define a multimodal SSMs. An approach similar to this was applied to multimodal magnetic resonance imaging (MRI) [93], but there are no examples of multimodal SSM models that have physics informed cross-layers.

Another part concerns training protocols. The first step will be to extend the task adapted reconstruction framework introduced by PI 1 in [4] to multichannel setting by using suitable DNN architectures from the above. Next is to adapt and apply techniques for pre-training and fine tuning against (small) domain specific well-curated datasets with self-supervision and model balancing [56]. These training protocols are commonly used for adapting pre-trained large language models (LLMs) to specialised domains [76, 53, 50, 57, 46] and they are also gaining traction in scientific machine learning (SciML) [79]. To further limit hallucinations, one can also consider the possibility to encode physics constraints with retrieval-augmented generation (RAG) techniques that were recently extended to computer vision [87, 85].

**Dynamic PET/CT.** Work here is led by PI 3 and its aim is to develop foundational models for reconstruction in dynamic PET/CT.

The main focus here will be on exploring computationally efficient and clinically viable ways of performing spatiotemporal PET/CT reconstruction from list-mode data. Temporal variation due to organ motion and tracer dynamics is here seen as a *deformation* with time dependent parameters. Optimisation schemes for the joint reconstruction of deformation and multi-channel image are well studied but they present two main challenges: they are computationally unfeasible and they rely on physical/biological models of deformation defined in image space. The latter limits the best achievable temporal resolution and also adds a computational cost due to the need of repeated transfer of the physical/biological information between the image and projection data domains (forward- and back-projection). PI 3 has previously showed that, feeding a DNN with sinograms patched in tokens corresponding to projection data pertaining to the same point or patch in image space [3, 49], improved both the denoising power and the generalisability of the model. The first step will therefore be to generalise this idea to the temporal domain. In short, we will feed suitable DNN:s with projection data,  $p(\underline{y}, t)$ , in patches corresponding to the trajectory  $\underline{x}(t)$  predicted by the relevant physical/biological deformation model and assess their reconstruction capability. In parallel, as an alternative approach, the effect of substituting simple convolution in temporal domain with physics informed DNNs will be explored. This is expected to achieve models that generalise well after training against small training data set. In order to address the heavy computational burden, we will select promising architectures from the work on foundation models to be used in this dual-domain reconstruction approach. A key point in what is proposed above is the access to representative training and test data sets. PIs 1 and 3 have previously developed methods for generating synthetic data sets that showed promising representativeness of physical motion in gated histogram projection data [16]. The intention is here to build on that work by including realistic simulations of line-of-response data and of biomarker kinematics. The line-of-response data will be simulated by random sampling



of forward-projected noisy images (note that, in forward projection here the temporal model of the deformation is included). Also, the technique described in [74] will be used to generate experimental dynamical phantoms in the PET/CT previously developed by PI 3. The above described work will be initially performed on problems of suitable size (e.g., corresponding to the PET/CT) and the last part of our efforts will be devoted to scaling up to clinically relevant cases. At this point, we will take advantage of the advancement produced in the foundational models part and we will have (at least) access to PET/CT data from around 200 patients collected by PI 3. It is relevant to mention that PI 3 is also involved in an initiative for the standardisation of PET data (ETSI), which could result in access to an even larger data set.

**Spectral CT:** Work here is led by PI 2 and it aims to develop foundational models for reconstruction in spectral CT. Spectral CT is a large-scale problem, e.g., a chest scan will typically generate 200 GB raw data. Hence, a key challenge is to manage memory footprint and computational burden in training and inference.

The first part is to establish a data platform and a digital twin for spectral CT. A key component is to generate a library of physiologically realistic concentration maps of 2–3 basis materials (multichannel phantoms) that can then be used to simulate dynamic, multispectral CT images. One approach is to inpaint realistic anatomy into existing dynamic phantoms, like XCAT [73]. This can be accomplished by style transfer trained on realistic static CT images.[17]

The second part is to adapt the foundation models for multichannel imaging developed by PI 1 to spectral CT. Here the idea is to extend the DNN architecture by PI 1 for extracting edge location/orientation [8] to a multichannel setting by introducing appropriate cross attention or cross-layer token fusion. These DNN architectures for geometric similarity can also be used in PET/CT. To further reduce risk of hallucinations, we complement the geometric similarity with physics models that couple the channels in spectral CT to each other.

The final part is to extend the above to a dynamical setting by leveraging on foundation models for motion that will be developed by PI 1. The key task is to enhance the aforementioned physics models to also accommodate different types of time-dependent attenuation such as rigid motion, tissue compression, and blood flow and perfusion. The resulting algorithms will be evaluated on simulated and clinical data, the latter obtained from a prototype photon-counting spectral CT scanner [7] in collaboration with Torkel Brismar and Staffan Holmin, professors in medical radiology and clinical neuroimaging at Karolinska Institutet (KI).

## Impact

The two use cases in medical imaging within the project have a societal impact and potential directly related to ‘Rich and Healthy Life’.

The installed base of photon-counting spectral CT scanners worldwide is growing rapidly. This offers a opportunity for fast clinical adoption of image reconstruction methods that account for spectral information and motion. This is especially the case in imaging iodine-enhancing tumours (cancer), haemorrhages (stroke), and blood vessel occlusions (cardiovascular diseases) where motion artefacts need to be accounted for in order to extract clinically important parameters. This will have important implications for patient health.

The clinical practice of PET/CT is, at present, limited to static PET that is semi-quantitative. Dynamic PET offers the advantage of being fully quantitative but it requires longer scan times along with invasive measurements (blood sampling). The proposed AI-tools will remove these two obstacles and enable dynamic PET as a routine clinical modality with a similar patient

throughput of static PET. Moreover, we expect our methods to provide better temporal and spatial resolution in PET based pre-clinical studies of drug kinematics.

Finally, the project will also make important contributions to developing DNN based approaches for multichannel imaging. This will require novel architectures that operate on the space of dynamical densities (higher-dimensional hyperobjects). Seen in a wider context, the approach is based on SciML [11] that blends handcrafted domain-specific models from applied mathematics with data-driven models from machine learning. As such, the primary scientific research theme for the project is Learn.

## Strategic relevance

The project relates directly to digitisation of image guided diagnostics and treatment. It also relates to gender equality as training and test data will be gender-wise unbiased. Finally, the project contributes to the [UN sustainable development goal 3](#) on good health and well-being, e.g., [indicator 3.4.1](#) (combating cancer, diabetes and cardiovascular and respiratory disease).


## Project team composition and resources

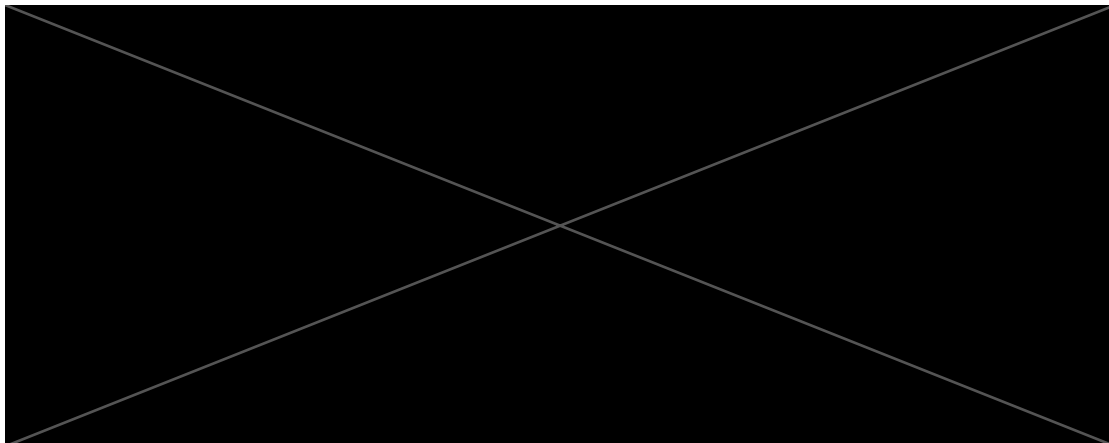
PI 1 has led development of physics informed DNNs for inverse modelling with applications to severely ill-posed image reconstruction problems. He will be responsible for the development of deep learning based techniques and associated mathematical theory. As part of [WASP](#) and [ELLIS](#), he also has access to compute infrastructure and research network for this task.

PI 2 has 14 years of research experience in photon-counting spectral CT. He will be responsible for the the spectral CT part of the project, including providing access to data, development of accurate physics models and supervision of a PhD student who will adapt the proposed image reconstruction techniques to photon-counting spectral CT. The algorithms will be tested on imaging data from a silicon-based photon-counting CT prototype by GE Healthcare available at the joint KTH/KI center [MedTechLabs](#).

PI 3 has comprehensive expertise from emission tomography, ranging from hardware (he built a [PET/CT](#)) and algorithms for image reconstruction. He will lead work on dynamic PET/CT in the project, and for generating and providing access to training and validation data sets.

## Budget

The requested funding of 3 MSEK for each PI will primarily cover salary costs for three PhDs and supervision the first three years. 



## Acronyms

<b>AI</b>	Artificial intelligence
<b>CT</b>	Computed tomography
<b>DNN</b>	Deep neural network
<b>KI</b>	Karolinska Institutet
<b>KTH</b>	KTH – Royal Institute of Technology
<b>LLM</b>	Large language model
<b>MRI</b>	Magnetic resonance imaging
<b>NLP</b>	Natural language processing
<b>PET</b>	Positron emission tomography
<b>PI</b>	Principal investigator
<b>RAG</b>	Retrieval-augmented generation
<b>SciML</b>	Scientific machine learning
<b>SOTA</b>	State-of-the-art
<b>SSM</b>	State space model

## References

- [1] Julián N. Acosta, Guido J. Falcone, Pranav Rajpurkar, and Eric J. Topol. “Multimodal biomedical AI”. In: *Nature Medicine* 28 (2022), pp. 1773–1784. DOI: [10.1038/s41591-022-01981-2](https://doi.org/10.1038/s41591-022-01981-2).
- [2] Anton Adelöw, Hamidreza Rashidy Kanan, and Massimiliano Colarieti-Tosti. “Transformer-based Sinusoidal Curve Learning for Sinogram Denoising”. In: *2024 IEEE Nuclear Science Symposium (NSS), Medical Imaging Conference (MIC) and Room Temperature Semiconductor Detector Conference (RTSD)*. 2024. DOI: [10.1109/NSS/MIC/RTSD57108.2024.10656409](https://doi.org/10.1109/NSS/MIC/RTSD57108.2024.10656409).
- [3] Anton Adelöw, Hamidreza Rashidy Kanan, Alessandro Guazzo, and Massimiliano Colarieti-Tosti. “Learned Primal Dual Reconstruction with Dual-Domain Transformers for PET”. In: *2024 IEEE Nuclear Science Symposium (NSS), Medical Imaging Conference (MIC) and Room Temperature Semiconductor Detector Conference (RTSD)*. 2024. DOI: [10.1109/NSS/MIC/RTSD57108.2024.10656509](https://doi.org/10.1109/NSS/MIC/RTSD57108.2024.10656509).
- [4] Jonas Adler, Sebastian Lunz, Olivier Verdier, Carola-Bibiane Schönlieb, and Ozan Öktem. “Task adapted reconstruction for inverse problems”. In: *Inverse Problems* 38.7 (2022), p. 075006. DOI: [10.1088/1361-6420/ac28ec](https://doi.org/10.1088/1361-6420/ac28ec).
- [5] Jonas Adler and Ozan Öktem. “Learned Primal-Dual Reconstruction”. In: *IEEE Transactions on Medical Imaging* 37.6 (2018), pp. 1322–1332. DOI: [10.1109/TMI.2018.2799231](https://doi.org/10.1109/TMI.2018.2799231).
- [6] Ameen Ali, Itamar Zimmerman, and Lior Wolf. “The Hidden Attention of Mamba Models”. In: *ArXiv e-prints* cs.LG.2403.01590 (2024). DOI: [10.48550/arXiv.2403.01590](https://doi.org/10.48550/arXiv.2403.01590).
- [7] Hakan Almqvist et al. “Initial Clinical Images From a Second-Generation Prototype Silicon-Based Photon-Counting Computed Tomography System”. In: *Academic Radiology* 31.2 (Nov. 2024), pp. 572–581. ISSN: 1076-6332. DOI: [10.1016/j.acra.2023.06.031](https://doi.org/10.1016/j.acra.2023.06.031).
- [8] Héctor Andrade-Loarca, Gitta Kutyniok, Ozan Öktem, and Philipp Petersen. “Extraction of digital wavefront sets using applied harmonic analysis and deep neural networks”. In: *SIAM Journal on Imaging Sciences* 12.4 (2019), pp. 1936–1966. DOI: [10.1137/19M1237594](https://doi.org/10.1137/19M1237594).
- [9] Hossein Arabi, Azadeh AkhavanAllaf, Amirhossein Sanaat, Isaac Shiri, and Habib Zaidi. “The promise of artificial intelligence and deep learning in PET and SPECT imaging”. In: *Physica Medica* 83 (2021), pp. 122–137. DOI: [10.1016/j.ejmp.2021.03.008](https://doi.org/10.1016/j.ejmp.2021.03.008).
- [10] Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. “Solving inverse problems using data-driven models”. In: *Acta Numerica* 28 (2019), pp. 1–174. DOI: [10.1017/S0962492919000059](https://doi.org/10.1017/S0962492919000059).
- [11] Nathan Baker et al. *Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence*. Tech. rep. U.S. Department of Energy Advanced Scientific Computing Research, 2019. DOI: [10.2172/1478744](https://doi.org/10.2172/1478744).
- [12] Vibha Balaji, Tzu-An Song, Masoud Malekzadeh, Pedram Heidari, and Joyita Dutta. “Artificial Intelligence for PET and SPECT Image Enhancement”. In: *Journal of Nuclear Medicine* 65.1 (2024), pp. 4–12. DOI: [10.2967/jnumed.122.265000](https://doi.org/10.2967/jnumed.122.265000).
- [13] Louis Blankemeier et al. “Merlin: A Vision Language Foundation Model for 3D Computed Tomography”. In: *ArXiv e-prints* cs.CV.2406.06512 (2024). DOI: [10.48550/arXiv.2406.06512](https://doi.org/10.48550/arXiv.2406.06512).



- [14] Davide Caffagni et al. “The Revolution of Multimodal Large Language Models: A Survey”. In: *Findings of the Association for Computational Linguistics: ACL 2024*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 13590–13618. DOI: [10.18653/v1/2024.findings-acl.807](https://doi.org/10.18653/v1/2024.findings-acl.807).
- [15] Richard E. Carson. “Tracer kinetic modeling in PET”. In: *Positron emission tomography: Basic sciences*. Ed. by Dale L. Bailey, David W. Townsend, Peter E. Valk, and Michael N. Maisey. Springer Verlag, 2005. Chap. 6, pp. 127–159. DOI: [10.1007/1-84628-007-9\\_6](https://doi.org/10.1007/1-84628-007-9_6).
- [16] Enza Cece, Pierre Meyrat, Enza Torino, Olivier Verdier, and Massimiliano Colarieti-Tosti. “Spatio-Temporal Positron Emission Tomography Reconstruction with Attenuation and Motion Correction”. In: *Journal of Imaging* 9.10 (2023), p. 231. DOI: [10.3390/jimaging9100231](https://doi.org/10.3390/jimaging9100231).
- [17] Yushi Chang, Kyle Lafata, William Paul Segars, Fang-Fang Yin, and Lei Ren. “Development of realistic multi-contrast textured XCAT (MT-XCAT) phantoms using a dual-discriminator conditional-generative adversarial network (D-CGAN)”. In: *Physics in Medicine & Biology* 65.6 (Mar. 2020), p. 065009. DOI: [10.1088/1361-6560/ab7309](https://doi.org/10.1088/1361-6560/ab7309).
- [18] Rahul Chanumolu, Likhita Alla, Pavankumar Chirala, Naveen Chand Chennampalli, and Bhanu Prakash Kolla. “Multimodal Medical Imaging Using Modern Deep Learning Approaches”. In: *2022 IEEE VLSI Device Circuit and System (VLSI DCS)*. 2022. DOI: [10.1109/VLSIDCS53788.2022.9811498](https://doi.org/10.1109/VLSIDCS53788.2022.9811498).
- [19] Chong Chen, Barbara Gris, and Ozan Öktem. “A new variational model for joint image reconstruction and motion estimation in spatiotemporal imaging”. In: *SIAM Journal on Imaging Sciences* 12.4 (2019), pp. 1686–1719. DOI: [10.1137/18M1234047](https://doi.org/10.1137/18M1234047).
- [20] Chong Chen and Ozan Öktem. “Indirect Image Registration with Large Diffeomorphic Deformations”. In: *SIAM Journal on Imaging* 11.1 (2018), pp. 575–617. DOI: [10.1137/17M113462](https://doi.org/10.1137/17M113462).
- [21] Zheyi Chen et al. “Evolution and Prospects of Foundation Models: From Large Language Models to Large Multimodal Models”. In: *Computers, Materials and Continua* 80.2 (2024), pp. 1753–1808. DOI: [10.32604/cmc.2024.052618](https://doi.org/10.32604/cmc.2024.052618).
- [22] Mats Danielsson, Mats Persson, and Martin Sjölin. “Photon-counting x-ray detectors for CT”. In: *Physics in Medicine & Biology* 66.3 (2021), 03TR01. DOI: [10.1088/1361-6560/abc5a5](https://doi.org/10.1088/1361-6560/abc5a5).
- [23] Sanuwani Dayarathna et al. “Deep learning based synthesis of MRI, CT and PET: Review and analysis”. In: *Medical Image Analysis* 92 (2024), p. 103046. DOI: [10.1016/j.media.2023.103046](https://doi.org/10.1016/j.media.2023.103046).
- [24] Junwei Duan, Jiaqi Xiong, Yinghui Li, and Weiping Ding. “Deep learning based multimodal biomedical data fusion: An overview and comparative review”. In: *Information Fusion* 112 (2024), p. 102536. DOI: [10.1016/j.inffus.2024.102536](https://doi.org/10.1016/j.inffus.2024.102536).
- [25] Alma Eguizabal, Ozan Öktem, and Mats Persson. “A deep learning one-step solution to material image reconstruction in photon counting spectral CT”. In: *SPIE Medical Imaging 2022: Physics of Medical Imaging*. Vol. 12031. 2022, pp. 239–243. DOI: [10.1117/12.2612426](https://doi.org/10.1117/12.2612426).
- [26] Alma Eguizabal, Mats Persson, and Ozan Öktem. “Material Decomposition for Photon Counting CT”. In: *16th International Meeting on Fully 3D Image Reconstruction in Radiology and Nuclear Medicine (Fully3D 2021)*. 2021. DOI: [10.48550/arXiv.2208.03360](https://doi.org/10.48550/arXiv.2208.03360).

- [27] Michael Elad, Bahjat Kwar, and Gregory Vaksman. “Image Denoising: The Deep Learning Revolution and Beyond – A Survey Paper”. In: *SIAM Journal on Imaging Sciences* 16.3 (2023), pp. 1594–1654. DOI: [10.1137/23M1545859](https://doi.org/10.1137/23M1545859).
- [28] Maryam Fallahpoor et al. “Deep learning techniques in PET/CT imaging: A comprehensive review from sinogram to image space”. In: *Computer Methods and Programs in Biomedicine* 243 (2024), p. 107880. DOI: [10.1016/j.cmpb.2023.107880](https://doi.org/10.1016/j.cmpb.2023.107880).
- [29] Nanyi Fei et al. “Towards artificial general intelligence via a multimodal foundation model”. In: *Nature Communications* 13 (2022), p. 3094. DOI: [10.1038/s41467-022-30761-2](https://doi.org/10.1038/s41467-022-30761-2).
- [30] Thomas Flohr et al. “Photon-counting CT review”. In: *Physica Medica: European Journal of Medical Physics* 79 (Nov. 2020), pp. 126–136. ISSN: 1120-1797. DOI: [10.1016/j.ejmp.2020.10.030](https://doi.org/10.1016/j.ejmp.2020.10.030).
- [31] Dustin A. Gress et al. “Ranking the Relative Importance of Image Quality Features in CT by Consensus Survey”. In: *Journal of the American College of Radiology* (2024). ISSN: 1546-1440. DOI: [10.1016/j.jacr.2024.10.006](https://doi.org/10.1016/j.jacr.2024.10.006).
- [32] Barbara Gris, Chong Chen, and Ozan Öktem. “Image reconstruction through metamorphosis”. In: *Inverse Problems* 36.2 (2020), p. 025001. DOI: [10.1088/1361-6420/ab5832](https://doi.org/10.1088/1361-6420/ab5832).
- [33] Albert Gu and Tri Dao. “Mamba: Linear-Time Sequence Modeling with Selective State Spaces”. In: *ArXiv e-prints cs.LG.2312.00752* (2023). DOI: [10.48550/arXiv.2312.00752](https://doi.org/10.48550/arXiv.2312.00752).
- [34] Alessandro Guazzo and Massimiliano Colarieti-Tosti. “Learned Primal Dual Reconstruction for PET”. In: *Journal of Imaging* 7.12 (2021), p. 248. DOI: [10.3390/jimaging7120248](https://doi.org/10.3390/jimaging7120248).
- [35] Zhe Guo, Xiang Li, Heng Huang, Ning Guo, and Quanzheng Li. “Deep Learning-Based Image Segmentation on Multimodal Medical Imaging”. In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 3.2 (2019), pp. 162–169. DOI: [10.1109/TRPMS.2018.2890359](https://doi.org/10.1109/TRPMS.2018.2890359).
- [36] Kai Han et al. “A Survey on Vision Transformer”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.1 (2023), pp. 87–110. DOI: [10.1109/TPAMI.2022.3152247](https://doi.org/10.1109/TPAMI.2022.3152247).
- [37] Fumio Hashimoto et al. “Deep learning-based PET image denoising and reconstruction: a review”. In: *Radiological Physics and Technology* 17 (2024), pp. 24–46. DOI: [10.1007/s12194-024-00780-3](https://doi.org/10.1007/s12194-024-00780-3).
- [38] Mohammed Hassanin, Saeed Anwar, Ibrahim Radwan, Fahad S. Khan, and Ajmal Mian. “Visual Attention Methods in Deep Learning: An In-Depth Survey”. In: *ArXiv e-prints cs.CV.2204.07756* (2024). DOI: [10.48550/arXiv.2204.07756](https://doi.org/10.48550/arXiv.2204.07756).
- [39] Andreas Hauptmann, Jonas Adler, Simon Arridge, and Ozan Öktem. “Multi-Scale Learned Iterative Reconstruction”. In: *IEEE Transactions on Computational Imaging* 8 (2020), pp. 843–856. DOI: [10.1109/TCI.2020.2990299](https://doi.org/10.1109/TCI.2020.2990299).
- [40] Andreas Hauptmann, Ozan Öktem, and Carola-Bibiane Schönlieb. “Image reconstruction in dynamic inverse problems with temporal models”. In: *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging*. Ed. by K. Chen, C.-B. Schönlieb, X. C. Tai, and L. Younes. Springer-Verlag, 2021. DOI: [10.1007/978-3-030-03009-4\\_83-1](https://doi.org/10.1007/978-3-030-03009-4_83-1).

- [41] Dennis Hein et al. “PPFM: Image Denoising in Photon-Counting CT Using Single-Step Posterior Sampling Poisson Flow Generative Models”. In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 8.7 (2024), pp. 788–799. DOI: [10.1109/TRPMS.2024.3410092](https://doi.org/10.1109/TRPMS.2024.3410092).
- [42] Thomas Wesley Holmes et al. “Ultrahigh-Resolution K-Edge Imaging of Coronary Arteries With Prototype Deep-Silicon Photon-Counting CT: Initial Results in Phantoms”. In: *Radiology* 311.3 (2024). PMID: 38916502, e231598. DOI: [10.1148/radiol.231598](https://doi.org/10.1148/radiol.231598).
- [43] Jiahao Huang et al. “Enhancing global sensitivity and uncertainty quantification in medical image reconstruction with Monte Carlo arbitrary-masked mamba”. In: *Medical Image Analysis* 99 (2025), p. 103334. DOI: [10.1016/j.media.2024.103334](https://doi.org/10.1016/j.media.2024.103334).
- [44] Ruihan Huang, Karin Larsson, and Mats U. Persson. “Deep-learning-based motion artifact reduction for photon-counting spectral cardiac CT”. In: *Proceedings of SPIE* 12925 (2024). Ed. by Rebecca Fahrig, John M. Sabol, and Ke Li, p. 1292507. DOI: [10.1117/12.3006362](https://doi.org/10.1117/12.3006362).
- [45] Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P. Lungren. “Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines”. In: *npj Digital Medicine* 3 (2020), Article no: 136. DOI: [10.1038/s41746-020-00341-z](https://doi.org/10.1038/s41746-020-00341-z).
- [46] Yixing Jiang et al. “Many-Shot In-Context Learning in Multimodal Foundation Models”. In: *ArXiv e-prints* cs.LG.2405.09798 (2024). DOI: [10.48550/arXiv.2405.09798](https://doi.org/10.48550/arXiv.2405.09798).
- [47] Peter Jörg. “Musiré: multimodal simulation and reconstruction framework for the radiological imaging sciences”. In: *Philosophical Transactions of the Royal Society A* 379.2204 (2021). DOI: [10.1098/rsta.2020.0190](https://doi.org/10.1098/rsta.2020.0190).
- [48] Hamidreza Rashidy Kanan, Anton Adelöw, and Massimiliano Colarieti-Tosti. “A Parallel Cross-Domain Reconstruction for PET”. In: *2024 IEEE Nuclear Science Symposium (NSS), Medical Imaging Conference (MIC) and Room Temperature Semiconductor Detector Conference (RTSD)*. 2024. DOI: [10.1109/NSS/MIC/RTSD57108.2024.10657117](https://doi.org/10.1109/NSS/MIC/RTSD57108.2024.10657117).
- [49] Hamidreza Rashidy Kanan, Anton Adelöw, and Massimiliano Colarieti-Tosti. “Cross-Domain Reconstruction Network Incorporating Sinogram Sinusoidal-Structure Transformer Denoiser and UNet for Low-Dose/Low-Count Sinograms”. In: *TechRxiv* (2024). DOI: [10.36227/techrxiv.172556974.48426599/v1](https://doi.org/10.36227/techrxiv.172556974.48426599/v1).
- [50] Haoran Lai et al. “E3D-GPT: Enhanced 3D Visual Foundation for Medical Vision-Language Model”. In: *ArXiv e-prints* eess.IV.2410.14200 (2024). DOI: [10.48550/arXiv.2410.14200](https://doi.org/10.48550/arXiv.2410.14200).
- [51] Lukas F. Lang, Sebastian Neumayer, Ozan Öktem, and Carola-Bibiane Schönlieb. “Template-Based Image Reconstruction from Sparse Tomographic Data”. In: *Applied Mathematics & Optimization* 82 (2020), pp. 1081–1109. DOI: [10.1007/s00245-019-09573-2](https://doi.org/10.1007/s00245-019-09573-2).
- [52] Johannes Leuschner et al. “Quantitative comparison of deep learning-based image reconstruction methods for low-dose and sparse-angle CT applications”. In: *Journal of Imaging* 7.3 (2021), Article no.: 44. DOI: [10.3390/jimaging7030044](https://doi.org/10.3390/jimaging7030044).
- [53] Chunyuan Li et al. “Multimodal Foundation Models: From Specialists to General-Purpose Assistants”. In: *ArXiv e-prints* cs.CV.2309.10020 (2023). DOI: [10.48550/arXiv.2309.10020](https://doi.org/10.48550/arXiv.2309.10020).

- [54] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. “Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions”. In: *ACM Computing Surveys* 56.10 (2024), 1–42 (Article No.: 264). DOI: [10.1145/3656580](https://doi.org/10.1145/3656580).
- [55] Leening P. Liu et al. “First-generation clinical dual-source photon-counting CT: ultra-low-dose quantitative spectral imaging”. In: *European Radiology* 32.12 (Dec. 2022), pp. 8579–8587. ISSN: 1432-1084. DOI: [10.1007/s00330-022-08933-x](https://doi.org/10.1007/s00330-022-08933-x).
- [56] Zihang Liu et al. “Model Balancing Helps Low-data Training and Fine-tuning”. In: *ArXiv e-prints* cs.LG.2410.12178 (2024). DOI: [10.48550/2410.12178](https://doi.org/10.48550/2410.12178).
- [57] Lingxiao Luo, Bingda Tang, Xuanzhong Chen, Rong Han, and Ting Chen. “VividMed: Vision Language Model with Versatile Visual Grounding for Medicine”. In: *ArXiv e-prints* cs.CV.2410.12694 (2024). DOI: [10.48550/arXiv.2410.12694](https://doi.org/10.48550/arXiv.2410.12694).
- [58] Keisuke Matsubara, Masanobu Ibaraki, Mitsutaka Nemoto, Hiroshi Watabe, and Yuichi Kimura. “A review on AI in PET imaging”. In: *Annals of Nuclear Medicine* 36 (2022), pp. 133–143. DOI: [10.1007/s12149-021-01710-8](https://doi.org/10.1007/s12149-021-01710-8).
- [59] Vishal Monga, Yuelong Li, and Yonina C. Eldar. “Algorithm Unrolling. Interpretable, Efficient Deep Learning for Signal and Image Processing”. In: *IEEE Signal Processing Magazine* 38.2 (2021), pp. 18–44. DOI: [10.1109/MSP.2020.3016905](https://doi.org/10.1109/MSP.2020.3016905).
- [60] Subhadip Mukherjee, Andreas Hauptmann, Ozan Öktem, Marcelo Pereyra, and Carola-Bibiane Schönlieb. “Learned Reconstruction Methods With Convergence Guarantees: A survey of concepts and applications”. In: *IEEE Signal Processing Magazine* 40.1 (2023), pp. 164–182. DOI: [10.1109/MSP.2022.3207451](https://doi.org/10.1109/MSP.2022.3207451).
- [61] Badri Narayana Patro and Vijay Srinivas Agneeswaran. “Mamba-360: Survey of State Space Models as Transformer Alternative for Long Sequence Modelling: Methods, Applications, and Challenges”. In: *ArXiv e-prints* cs.LG.2404.16112 (2024). DOI: [10.48550/arXiv.2404.16112](https://doi.org/10.48550/arXiv.2404.16112).
- [62] Mats Persson, Adam Wang, and Norbert J. Pelc. “Detective quantum efficiency of photon-counting CdTe and Si detectors for computed tomography: a simulation study”. In: *Journal of Medical Imaging* 7.4 (2020), p. 043501. DOI: [10.1117/1.JMI.7.4.043501](https://doi.org/10.1117/1.JMI.7.4.043501).
- [63] Mats Persson et al. “Energy-resolved CT imaging with a photon-counting silicon-strip detector”. In: *Physics in Medicine & Biology* 59.22 (2014), p. 6709. DOI: [10.1088/0022-3727/59/22/6709](https://doi.org/10.1088/0022-3727/59/22/6709).
- [64] Michael E. Phelps, ed. *PET: Molecular Imaging and Its Biological Applications*. Springer Verlag, 2004. DOI: [10.1007/978-0-387-22529-6](https://doi.org/10.1007/978-0-387-22529-6).
- [65] Michael E. Phelps, Edward J. Hoffman, Nizar A. Mullani, and Michel M. Ter-Pogossian. “Application of annihilation coincidence detection to transaxial reconstruction tomography”. In: *Journal of nuclear medicine* 16.3 (1975), pp. 210–224.
- [66] Camille Pouchol, Olivier Verdier, and Ozan Öktem. “Spatiotemporal PET reconstruction using ML-EM with learned diffeomorphic deformation”. In: *2nd International Workshop on Machine Learning for Medical Image Reconstruction (MLMIR 2019)*. Ed. by F. Knoll, A. Maier, D. Rueckert, and J. C. Ye. Vol. 11905. Lecture Notes in Computer Science. Springer-Verlag, 2019, pp. 151–162. DOI: [10.1007/978-3-030-33843-5\\_14](https://doi.org/10.1007/978-3-030-33843-5_14).
- [67] Kishore Rajendran et al. “First Clinical Photon-counting Detector CT System: Technical Evaluation”. In: *Radiology* 303.1 (2022), pp. 130–138. DOI: [10.1148/radiol.212579](https://doi.org/10.1148/radiol.212579).

- [68] Andrew J. Reader et al. “Deep Learning for PET Image Reconstruction”. In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 5.1 (2020), pp. 1–25. DOI: [10.1109/TRPMS.2020.3014786](https://doi.org/10.1109/TRPMS.2020.3014786).
- [69] Jevgenija Rudzusika, Buda Bajić, Thomas Koehler, and Ozan Öktem. “3D helical CT Reconstruction with a Memory Efficient Learned Primal-Dual Architecture”. In: *IEEE Transactions on Computational Imaging* 10 (2024), pp. 1414–1424. DOI: [10.1109/TCI.2024.3463485](https://doi.org/10.1109/TCI.2024.3463485).
- [70] Aria M. Salyapongse et al. “Spatial Resolution Fidelity Comparison Between Energy Integrating and Deep Silicon Photon Counting CT: Implications for Pulmonary Imaging”. In: *Journal of Thoracic Imaging* 39.6 (2024). DOI: [10.1097/RTI.0000000000000788](https://doi.org/10.1097/RTI.0000000000000788).
- [71] Veit Sandfort et al. “Spectral photon-counting CT in cardiovascular imaging”. In: *Journal of Cardiovascular Computed Tomography* 15.3 (2021), pp. 218–225. DOI: [10.1016/j.jcct.2020.12.005](https://doi.org/10.1016/j.jcct.2020.12.005).
- [72] Daan Schouten et al. “Navigating the landscape of multimodal AI in medicine: a scoping review on technical challenges and clinical applications”. In: *ArXiv e-prints* cs.AI.2411.03782 (2024). DOI: [10.48550/arXiv.2411.03782](https://doi.org/10.48550/arXiv.2411.03782).
- [73] Paul W. Segars, Gregory M. Sturgeon, Marc S. Mendonca, Jason Grimes, and Benjamin M. W. Tsui. “4D XCAT phantom for multimodality imaging research”. In: *Medical Physics* 37.9 (2010), pp. 4902–4915. DOI: [10.1118/1.3480985](https://doi.org/10.1118/1.3480985).
- [74] Ekaterina Shanina, Benjamin A. Spencer, Tiantian Li, Jinyi Qi, and Simon R. Cherry. “A universal activity painting phantom for positron emission tomography”. In: *2024 IEEE Nuclear Science Symposium (NSS), Medical Imaging Conference (MIC) and Room Temperature Semiconductor Detector Conference (RTSD)*. IEEE. 2024, pp. 1–1. DOI: [10.1109/NSS/MIC/RTSD57108.2024.10657713](https://doi.org/10.1109/NSS/MIC/RTSD57108.2024.10657713).
- [75] Hui Shen, Zhongwei Wan, Xin Wang, and Mi Zhang. “Famba-V: Fast Vision Mamba with Cross-Layer Token Fusion”. In: *ArXiv e-prints* cs.CV.2409.09808 (2024). DOI: [10.48550/arXiv.2409.09808](https://doi.org/10.48550/arXiv.2409.09808).
- [76] Prashant Shrestha, Sanskar Amgain, Bidur Khanal, Cristian A. Linte, and Binod Bhattarai. “Medical Vision Language Pretraining: A survey”. In: *ArXiv e-prints* cs.CV.2312.06224 (2023). DOI: [10.48550/arXiv.2312.06224](https://doi.org/10.48550/arXiv.2312.06224).
- [77] Joakim da Silva et al. “Resolution characterization of a silicon-based, photon-counting computed tomography prototype capable of patient scanning”. In: *Journal of Medical Imaging* 6.4 (2019), pp. 1–9. DOI: [10.1117/1.JMI.6.4.043502](https://doi.org/10.1117/1.JMI.6.4.043502).
- [78] Emanuel Ström, Mats Persson, Alma Eguizabal, and Ozan Öktem. “Photon-Counting CT Reconstruction with a Learned Forward Operator”. In: *IEEE Transactions on Computational Imaging* 8 (2022), pp. 536–550. DOI: [10.1109/TCI.2022.3183405](https://doi.org/10.1109/TCI.2022.3183405).
- [79] Shashank Subramanian et al. “Towards foundation models for scientific machine learning: characterizing scaling and transfer behavior”. In: *Advances in Neural Information Processing Systems* 37 (*NeurIPS 2023*). 2023, 71242–71262 (article number: 3119). URL: <https://dl.acm.org/doi/10.5555/3666122.3669241>.
- [80] Rolf Symons et al. “Dual-contrast agent photon-counting computed tomography of the heart: initial experience”. In: *The International Journal of Cardiovascular Imaging* 33.8 (2017), pp. 1253–1261. DOI: [10.1007/s10554-017-1104-4](https://doi.org/10.1007/s10554-017-1104-4).



- [81] Rolf Symons et al. “Feasibility of Dose-reduced Chest CT with Photon-counting Detectors: Initial Results in Humans”. In: *Radiology* 285.3 (2017), pp. 980–989. DOI: [10.1148/radiol.2017162587](https://doi.org/10.1148/radiol.2017162587).
- [82] Rolf Symons et al. “Photon-counting CT for simultaneous imaging of multiple contrast agents in the abdomen: An in vivo study”. In: *Medical Physics* 44.10 (2017), pp. 5120–5127. DOI: [10.1002/mp.12301](https://doi.org/10.1002/mp.12301).
- [83] Mustafa Takesh. “The potential benefit by application of kinetic analysis of PET in the clinical oncology”. In: *International Scholarly Research Notices* 2012.1 (2012), Article ID: 349351. DOI: [10.5402/2012/349351](https://doi.org/10.5402/2012/349351).
- [84] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. “Multimodal Large Language Models: A Survey”. In: *2023 IEEE International Conference on Big Data*. 2024. DOI: [10.1109/BigData59044.2023.10386743](https://doi.org/10.1109/BigData59044.2023.10386743).
- [85] Peng Xia et al. “MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models”. In: *ArXiv e-prints* cs.LG.2410.13085 (2024). DOI: [10.48550/arXiv.2410.13085](https://doi.org/10.48550/arXiv.2410.13085).
- [86] Chaodong Xiao, Minghan Li, Zhengqiang Zhang, Deyu Meng, and Lei Zhang. “Spatial-Mamba: Effective Visual State Space Models via Structure-Aware State Fusion”. In: *ArXiv e-prints* cs.CV.2410.15091 (2024). DOI: [10.48550/arXiv.2410.15091](https://doi.org/10.48550/arXiv.2410.15091).
- [87] Shi Yu et al. “VisRAG: Vision-based Retrieval-augmented Generation on Multi-modality Documents”. In: *ArXiv e-prints* cs.IR.2410.10594 (2024). DOI: [10.48550/arXiv.2410.10594](https://doi.org/10.48550/arXiv.2410.10594).
- [88] Vicky Zayats, Peter Chen, Melissa Ferrari, and Dirk Padfield. “Zipper. A Multi-Tower Decoder Architecture for Fusing Modalities”. In: *ArXiv e-prints* cs.LG.2405.18669 (2024). DOI: [10.48550/arXiv.2405.18669](https://doi.org/10.48550/arXiv.2405.18669).
- [89] Fei Zhao, Chengcui Zhang, and Baocheng Geng. “Deep Multimodal Data Fusion”. In: *ACM Computing Surveys* 56.9 (2024), 1–36 (Article No.: 2016). DOI: [10.1145/3649447](https://doi.org/10.1145/3649447).
- [90] Wayne Xin Zhao et al. “A Survey of Large Language Models”. In: *ArXiv e-prints* cs.CL.2303.18223 (2024). DOI: [10.48550/arXiv.2303.18223](https://doi.org/10.48550/arXiv.2303.18223).
- [91] Xianbing Zhao, Soujanya Poria, Xuejiao Li, Yixin Chen, and Buzhou Tang. “Toward Robust Multimodal Learning using Multimodal Foundational Models”. In: *ArXiv e-prints* cs.CV.2401.13697 (2024). DOI: [10.48550/arXiv.2401.13697](https://doi.org/10.48550/arXiv.2401.13697).
- [92] Ye Zhu, Yu Wu, Nicu Sebe, and Yan Yan. “Vision+X: A Survey on Multimodal Learning in the Light of Data”. In: *ArXiv e-prints* cs.CV.2210.02884 (2024). DOI: [10.48550/arXiv.2210.02884](https://doi.org/10.48550/arXiv.2210.02884).
- [93] Jing Zou et al. “MMR-Mamba: Multi-Modal MRI Reconstruction with Mamba and Spatial-Frequency Information Fusion”. In: *ArXiv e-prints* eess.IV.2406.18950 (2024). DOI: [10.48550/arXiv.2406.18950](https://doi.org/10.48550/arXiv.2406.18950).