Dual-task kidney MR segmentation with Transformers in autosomal-dominant polycystic kidney disease

Pierre-Henri Conze, Gustavo Andrade-Miranda, Yannick Lemeur, Emilie Cornec-Le Gall, François Rousseau

Computerized Medical Imaging and Graphics 113 (2024) 102349

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Dual-task kidney MR segmentation with transformers in autosomal-dominant polycystic kidney disease

Pierre-Henri Conze [a,b,*], Gustavo Andrade-Miranda [b], Yannick Le Meur [c,d], Emilie Cornec-Le Gall [c,e], François Rousseau [a,b]

[a] IMT Atlantique, LaTIM UMR 1101, Technopôle Brest-Iroise, 29238 Brest, France
[b] LaTIM UMR 1101, Inserm, IBRBS, 22 rue Camille Desmoulins, 29200 Brest, France
[c] Department of Nephrology, University Hospital of Brest, bd Tanguy Prigent, 29200 Brest, France
[d] LBAI UMR 1227, Inserm, 9 rue Félix le Dantec, 29200 Brest, France
[e] UMR 1078, Inserm, IBRBS, 22 rue Camille Desmoulins, 29238 Brest, France

ARTICLE INFO

Keywords:
Kidney segmentation
Semantic segmentation
Multi-task learning
Transformers
Polycystic kidney disease

ABSTRACT

Autosomal-dominant polycystic kidney disease is a prevalent genetic disorder characterized by the development of renal cysts, leading to kidney enlargement and renal failure. Accurate measurement of total kidney volume through polycystic kidney segmentation is crucial to assess disease severity, predict progression and evaluate treatment effects. Traditional manual segmentation suffers from intra- and inter-expert variability, prompting the exploration of automated approaches. In recent years, convolutional neural networks have been employed for polycystic kidney segmentation from magnetic resonance images. However, the use of Transformer-based models, which have shown remarkable performance in a wide range of computer vision and medical image analysis tasks, remains unexplored in this area. With their self-attention mechanism, Transformers excel in capturing global context information, which is crucial for accurate organ delineations. In this paper, we evaluate and compare various convolutional-based, Transformers-based, and hybrid convolutional/Transformers-based networks for polycystic kidney segmentation. Additionally, we propose a dual-task learning scheme, where a common feature extractor is followed by per-kidney decoders, towards better generalizability and efficiency. We extensively evaluate various architectures and learning schemes on a heterogeneous magnetic resonance imaging dataset collected from 112 patients with polycystic kidney disease. Our results highlight the effectiveness of Transformer-based models for polycystic kidney segmentation and the relevancy of exploiting dual-task learning to improve segmentation accuracy and mitigate data scarcity issues. A promising ability in accurately delineating polycystic kidneys is especially shown in the presence of heterogeneous cyst distributions and adjacent cyst-containing organs. This work contribute to the advancement of reliable delineation methods in nephrology, paving the way for a broad spectrum of clinical applications.

1. Introduction

Autosomal-dominant polycystic kidney disease (ADPKD) is a systemic genetic disorder that is the most common hereditary renal disease, affecting around 12.5 million people worldwide (Chapman et al., 2015). This pathology is characterized by the enlargement of kidneys due to the progressive development of renal cysts. Fourth leading cause of kidney failure, it requires dialysis or kidney transplantation for the majority of patients (Cornec-Le Gall et al., 2019). ADPKD, whose degree of phenotypic variability among affected individuals is extremely broad, can also manifest with extra-renal symptoms such as the presence of cysts in the liver, intra-cranial aneurysms or cardiac valvular disease. ADPKD is most commonly due to mutations in the *PKD1* or *PKD2* genes, respectively in 78% and 15% of disease pedigrees (Cornec-Le Gall et al., 2019). Patients with *PKD1* truncating (*PKD1t*) or non-truncating (*PKD1nt*) mutations have more severe disease with larger kidneys and a larger number of cysts compared to patients with *PKD2* mutations or without detected mutation (Cornec-Le Gall et al., 2018). The continuous growth of cysts in ADPKD leads to a progressive increase in total kidney volume (TKV). TKV is the most important imaging biomarker for quantifying the severity of ADPKD. As recognized by the US Food and Drug Administration and the European Medicine Agency, it is a prognostic enrichment biomarker in patients with ADPKD that predicts future renal function decline (Higashihara

* Corresponding author at: IMT Atlantique, LaTIM UMR 1101, Technopôle Brest-Iroise, 29238 Brest, France.
E-mail address: pierre-henri.conze@imt-atlantique.fr (P.-H. Conze).

https://doi.org/10.1016/j.compmedimag.2024.102349
Received 23 June 2023; Received in revised form 21 January 2024; Accepted 1 February 2024

et al., 2014). In early disease, additional value over renal function measurements can stay within normal ranges for a prolonged period of time due to hyper-filtration of remaining nephrons (Grantham et al., 2006). TKV is used in clinical care to assess the risk of individual disease progression and select patients with rapid progression for Tolvaptan treatment or clinical trials (van Gastel et al., 2019). TKV is also often employed as primary or secondary end-point to assess treatment effects (van Gastel et al., 2019).

Reaching TKV measurements requires to perform polycystic kidney segmentation which is usually done manually. Due to its superior soft tissue contrast, accuracy, and non-ionizing radiation, magnetic resonance (MR) imaging is the imaging modality of choice in this context. However, manual delineation is prone to a strong intra- and inter-expert variability due to severe alterations in the morphology, non-uniform cyst formation as well as the presence of adjacent liver cysts. As a consequence, several methods have been proposed to perform TKV computation in a simple manner through ellipsoid volume equations (Higashihara et al., 2015). To provide a more reliable estimation of TKV, particularly in the context of clinical trials, various image processing techniques have been proposed including random walks (Daum et al., 2007) or seeded region-growing (Mignani et al., 2011). However, these methods generally tend to produce sub-optimal results and may require manual adjustments. Alternative model-based approaches such as level set (Kim et al., 2016) or active contours (Racimora et al., 2010) can enable the isolation of the kidney areas from the surrounding abdominal anatomy. These models can be optimized by adjusting their parameters or iteratively obtained through a differential equation which is guided by the image properties and expected kidney shape. Nevertheless, a major drawback of such approaches is the need for prior knowledge (e.g. shape, texture). Therefore, model-based approaches may not be flexible enough to adapt to variations in MR scans. Moreover, in ADPKD, cysts may strongly vary in shape, size, and texture, making it difficult to define a single mathematical model that encompasses all cysts. This motivated the development of more sophisticated, reproducible, and operator-independent delineation strategies.

With the rise of deep learning which enables to design segmentation pipelines without defining hand-crafted features, the polycystic kidney segmentation and derived TKV computation tasks have been gradually automatized and improved using convolutional neural networks (CNN). While extensive research has been conducted using deep neural networks for tumoral kidney segmentation from computed tomography (CT) scans (Heller et al., 2021) and apart from the study of healthy kidney delineation for MR images (Kavur et al., 2021; Conze et al., 2021), the use of MR imaging remains relatively limited, especially for ADPDK patients. In particular, Kline et al. developed in Kline et al. (2017) a U-shaped architecture (Ronneberger et al., 2015) comprising five downsampling and upsampling blocks, along with skip connections. Subsequently, 11 instances of this architecture were employed to create an artificial multi-observer deep neural network, which was trained on distinct data subsets for the automated segmentation of polycystic kidneys from MR scans. The final segmentation was obtained by simulating a multi-observer majority voting scheme. However, despite the promising performance, the computational complexity and time required for the segmentation process are high, which may limit the feasibility in practical applications. Bevilacqua et al. conducted in Bevilacqua et al. (2019) a comparison between the 2D SegNet segmentation network (Badrinarayanan et al., 2017) and a two-step classification approach involving region of interest (ROI) detection using R-CNN, followed by semantic segmentation of the extracted ROIs. Interestingly, the authors found that performing segmentation on the entire MR volume was more reliable than on extracted ROIs. More recently, Guo et al. presented in Guo et al. (2022) a cascaded CNN including two 2D UNet models based on the ResNet34 backbone. Both networks were trained independently with three slices as inputs and the ground truth as additional input for the second network. During inference, delineation masks for all slices from a given subject were predicted in a sequential manner. Nevertheless, the manual selection of subjects for training was shown to introduce variability in the delineation results. Furthermore, the algorithm failed to discriminate renal parenchyma from small renal cysts. In line with previous works, Goel et al. developed in Goel et al. (2022) a deep architecture extending UNet with EfficientNet (Tan and Le, 2019) as encoder. Despite accurate kidney contours, notable errors were reported in the following circumstances: fluid-filled stomach, distended urinary bladder, hemorrhagic renal cysts and hepatic cysts in the vicinity of the right kidney. In contrast to prior studies, Raj et al. introduced in Raj et al. (2022) novel improvements to the UNet architecture (Ronneberger et al., 2015) for image segmentation. Specifically, they incorporated three attention mechanisms into the UNet framework: convolutional block attention, squeeze and excitation attention, and channel attention. These mechanisms aimed at enhancing the ability to focus on salient features, thereby improving the accuracy of the segmentation task. Additionally, the authors implemented a cosine loss function, which has been shown to be effective for training deep models on small datasets. To further enhance the generalizability of their network, they applied a sharpness-aware minimization technique, which regularizes the network by penalizing predictions that are too sharp or too blurred. However, the potential heterogeneous distributions of cysts throughout the abdomen still poses challenges in accurately distinguishing polycystic kidneys from surrounding organs. As a result, models may occasionally over-segment the kidneys by erroneously delineating parts of other abdominal structures that also contain cysts, such as the liver.

Polycystic kidney delineation from MR data has predominantly relied on CNN architectures so far (Zöllner et al., 2021). However, recent developments in medical image analysis have demonstrated the potential of Transformer-based models, which have shown superior performance in various computer vision applications (Dosovitskiy et al., 2020; Carion et al., 2020; Touvron et al., 2021). Vision Transformer (ViT) models have especially gained significant attention in medical image analysis tasks, including medical image segmentation, with an exponential growth of related publications (Jun et al., 2021; Shamshad et al., 2022). Unlike CNNs, Transformers do not require any convolution or pooling operations but instead rely on self-attention mechanisms to model the relationships between different image regions. This approach has shown to be particularly effective for capturing global context information in medical images, which can be critical to reach sufficiently efficient delineations that meet clinical requirements. Recently, novel Transformer-based models such as UNETR (Hatamizadeh et al., 2022b) and Swin UNETR (Hatamizadeh et al., 2022a) have been proposed and have shown promising semantic segmentation results. Both Swin UNETR and UNETR maintain the encoder–decoder UNet architecture but differ in the encoder component. Specifically, UNETR leverages a 3D ViT as encoder which reshapes the last feature map and upsamples it before using CNN upsampling with multi-level feature aggregation to generate segmentation outputs. Despite delivering competitive results, the vanilla ViT suffers from shortcomings when making dense predictions due to the absence of prior information from images. Hierarchical ViT models such as Swin UNETR (Hatamizadeh et al., 2022a) address this issue by injecting more-specific inductive biases derived from CNN-like features into the Transformers. Thus, Swin UNETR computes local attention through shifted windows, starting with small-sized patches and progressively merging neighboring patches in the subsequent layers. Although Transformer-based models have gained significant interest and have shown promising results in various applications (Cirrincione et al., 2023; Dhamija et al., 2023; Andrade-Miranda et al., 2023), their use in polycystic kidney segmentation from MR images has not been investigated to our knowledge.

In addition to model architecture, multi-task learning has also become increasingly important in medical image analysis (Zhao et al., 2022; Conze et al., 2023). The integration of multi-task learning offers several advantages. First, it avoids redundant learning of common-shared features for different tasks, leading to a substantial reduction

in overall memory consumption. Second, it has the capacity to learn more generalized features by averaging inherent noise patterns among various tasks. Third, it can prioritize crucial features that are usually challenging to distinguish within a single-task framework. Lastly, it introduces inductive biases to mitigate the overfitting problem, proving superior to conventional regularization methods. Multi-task learning has been used for automatic breast mass detection (Yan et al., 2021) through a unified Siamese network leveraging craniocaudal (CC) and mediolateral-oblique (MLO) views. This network combines patch-level mass/non-mass classification with dual-view mass matching to fully exploit multi-view information. In prostate cancer, Duran et al. (2022) proposed an all-in-one multi-class network. Using a parallel and cascaded approach for multi-task learning, it encodes MR information into a latent space. Two decoding branches follow: the first for binary prostate segmentation and the second using the segmented prostate as a prior for lesion detection and grading through an attention mechanism. The prostate decoder's output serves as a soft attention map for the lesion decoder, enhancing its performance. A multi-task, multi-domain bone MR segmentation method was proposed in Boutillon et al. (2022) to address the scarcity of pediatric imaging datasets by simultaneously considering multiple intensity domains and segmentation tasks, leveraging shared features across imaging datasets. To enhance generalization, a transfer learning scheme from natural image classification was applied. Additionally, a multi-scale contrastive regularization was used to encourage domain-specific clusters in shared representations, and multi-joint anatomical priors were incorporated to ensure anatomically consistent predictions. Despite the significance of multi-task learning in medical imaging, its potential benefits remain largely unexplored in the context of polycystic kidney segmentation, creating a notable gap in current studies. Integrating multi-task learning has the potential to improve the model's generalizability and reduce overfitting. Furthermore, joint training of multiple tasks can mitigate issues related to data scarcity, while shared network parameters allow models to acquire a more efficient and compact representation of the data, especially when tasks are inter-related or share commonalities.

With the aim of designing the best possible polycystic kidney delineation system from MR scans, our contributions are three folds. First, this paper aims at evaluating and comparing various purely CNN-based, Transformers-based, and hybrid CNN/Transformers-based networks in the context of polycystic kidney segmentation. Second, we propose to extend these backbones with a simple yet effective dual-task learning scheme involving a common feature extractor followed by per-kidney decoders. Third, a comprehensive evaluation is provided on a heterogeneous MR imaging dataset collected from 111 patients with ADPDK. The structure of this paper is as follows. In Section 2, we introduce three types of networks (i.e. CNN-based, Transformers-based, and hybrid CNN/Transformers-based) and extend them through dual-task learning. In Section 3, we provide a detailed description of the implementation and evaluation strategy. The obtained results are presented and discussed in Section 4. Finally, Section 5 summarizes the key findings of our study and provides insights into future research directions.

2. Method

2.1. Problem formulation

Let $\boldsymbol{I} \in \mathbb{R}^{H\times W\times D}$ denote a input 3D MR volume with dimensions $H\times W\times D$. Each 2D slice $\boldsymbol{x} \in \mathbb{R}^{H\times W}$ from $\boldsymbol{I}$ is associated with a ground truth annotation mask $\boldsymbol{y} \in [0,1]^{H\times W\times C}$ where C is the number of classes. The 2D segmentation problem can be formulated as finding the function $\phi : \boldsymbol{x} \rightarrow \phi(\boldsymbol{x};\boldsymbol{\Theta}) = \hat{\boldsymbol{y}}$ with weights $\boldsymbol{\Theta}$ that produce the best mapping between the slice $\boldsymbol{x}$ and the label map $\boldsymbol{y}$ from N training samples, by optimizing a loss function $\mathcal{L}_{\phi}(\boldsymbol{y},\hat{\boldsymbol{y}})$. It is now common practice to formulate the ϕ function as a U-shaped encoder–decoder network. The encoder, denoted by $f(\cdot)$, takes the input $\boldsymbol{x}$ and compresses it into a hidden representation $\boldsymbol{h} = f(\boldsymbol{x})$. The decoder, denoted by $g(\cdot)$, takes the compressed representation $\boldsymbol{h}$ as input and generates the final predicted segmentation map, $\hat{\boldsymbol{y}} = g(\boldsymbol{h})$.

In multi-task segmentation, each $\boldsymbol{x}$ sample is associated with a set of masks $\boldsymbol{Y} = \{\boldsymbol{y}_i\}$, where $\boldsymbol{y}_i$ is the ground truth delineation for the ith task. A common approach in this context is to use a global encoder to extract features from the input image for all tasks, followed by individual decoder branches (i.e. one decoder for each task). This enables the network to share a common feature extraction process while also being able to learn task-specific features. The global encoder $f(\cdot)$ extracts features from the input image, with the resulting hidden representation $\boldsymbol{h}$ serving as input for each of the task-specific decoders. The task-specific decoder for the ith task, denoted as $g_i(\cdot)$, is responsible for generating the segmentation map $\hat{\boldsymbol{y}}_i = g_i(\boldsymbol{h})$. Finally, the function $\phi(\boldsymbol{x};\boldsymbol{\Theta})$ that produces the best mapping between $\boldsymbol{x}$ and $\boldsymbol{Y}$ is obtained by optimizing the following joint loss function:

$$\mathcal{L}_{\phi}(\boldsymbol{Y},\hat{\boldsymbol{Y}}) = \sum_{i=1}^{K} \lambda_i \mathcal{L}_{\phi i}(\boldsymbol{y}_i,\hat{\boldsymbol{y}}_i) \tag{1}$$

where K is the number of segmentation tasks, $\hat{\boldsymbol{Y}}$ the set of k predicted segmentation maps, $\mathcal{L}_{\phi i}(\boldsymbol{y}_i,\hat{\boldsymbol{y}}_i)$ the individual loss function for the ith task and λ_i an hyper-parameter that controls the relative importance of the ith task with respect to the remaining ones, with $\sum_{i=1}^{K}\lambda_i = 1$.

2.2. Learning schemes

Segmenting both left and right kidneys for patients with ADPDK is challenging because each of them are present in a different spatial context, close to different anatomical structures. For instance, the right kidney is in interaction with the liver whereas the left kidney is in the vicinity of the spleen. In addition, the spatial distribution and heterogeneity of cysts can greatly vary from one kidney to another. In this context, three learning schemes can be considered (Fig. 1) towards polycystic kidney segmentation with deep learning:

- both organs (BO): the most common configuration consists in exploiting a single deep network segmenting both left and right kidneys, without any distinction between them. Therefore, the network made of one single encoder $f_{\mathrm{BK}}(\cdot)$ followed by one single decoder $g_{\mathrm{BK}}(\cdot)$ performs a binary segmentation task, distinguishing between renal (including cysts) and non-renal tissues, whatever the laterality.
- independent (IND): this strategy involves two separate encoder–decoder networks : $f_{\mathrm{LK}}(\cdot)$ followed by $g_{\mathrm{LK}}(\cdot)$ and $f_{\mathrm{RK}}(\cdot)$ followed by $g_{\mathrm{RK}}(\cdot)$. Each of them performs a binary segmentation task without any weight sharing. One aims at delineating the left kidney whereas the second segment the right kidney.
- dual-task (DT): this multi-task scheme makes use of a single network comprising one single encoder $f_{\mathrm{LK+RK}}(\cdot)$ and two task-specific decoders $g_{\mathrm{LK}}(\cdot)$ and $g_{\mathrm{RK}}(\cdot)$, one for each kidney. Features arising from the encoder are common to both task.

In IND and DT configurations, both results are then fused through a simple union operator to get a full renal cartography, comprising both left and right kidneys.

2.3. Network architectures

In this work, we explore the use of different deep models for polycystic kidney segmentation: CNN-based (v19pUNet (Conze et al., 2020)), hybrid CNN/Transformer-based (TransUNet (Chen et al., 2021), MedT (Valanarasu et al., 2021), SwinUNetV2 (Liu et al., 2022)) and Transformer-based (Segmenter (Strudel et al., 2021)) . Each architecture, explained in detail below, is employed in a single- (BO, IND) and dual-task (DT) learning fashion.

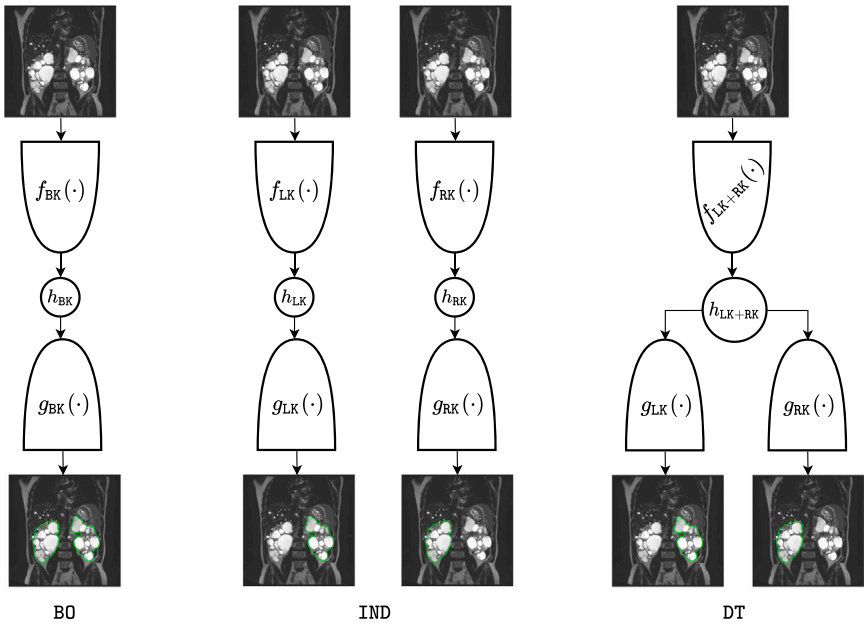

Fig. 1. BO, IND and DT learning schemes for kidney segmentation in patients with ADPKD. For sake of clarity, skip-connections are not displayed. Refer to text for further details.

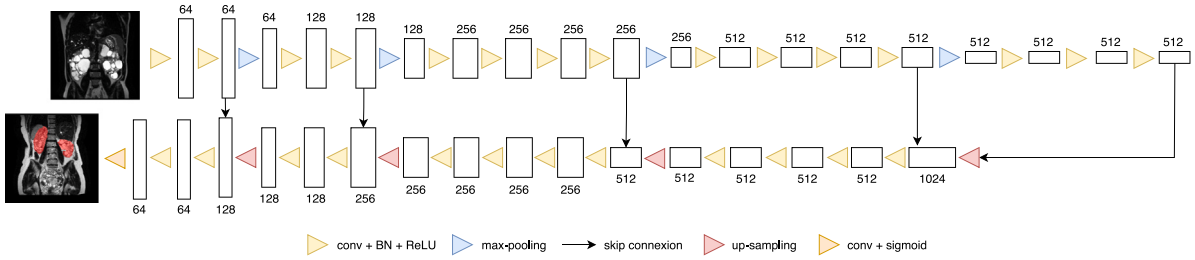

Fig. 2. v19pUNet (Conze et al., 2021), CNN-based encoder–decoder architecture.

v19pUNet. The v19pUNet model is a CNN-based architecture successfully applied to the CHAOS challenge (Conze et al., 2020; Kavur et al., 2021), built upon the standard UNet (Ronneberger et al., 2015) and incorporating a VGG19 backbone (Simonyan and Zisserman, 2014) as encoder (Fig. 2). Compared to UNet, the first convolutional layer of v19UNet generates 64 channels instead of 32, and the channel count doubles after each max pooling operation, until it reaches 512 (256 only for UNet). Additionally, after the second max pooling, v19pUNet has four consecutive layers per pattern, unlike UNet's two consecutive layers as used in Ronneberger et al. (2015). This VGG19-like encoder branch is pre-trained on ImageNet (Russakovsky et al., 2015) to benefit from 1 million non-medical data collected for object recognition purposes. While VGG19 includes fully-connected and softmax top layers, these are omitted in v19pUNet. Instead, the last three convolutional layers of VGG-19 are used as a central part to separate both contracting and expanding paths. To get a symmetric architecture, the decoder branch is extended in the same way as the encoder by adding four convolutional layers and more feature channels. At the end, a final 1×1 convolutional layer with a sigmoid activation achieves pixel-wise segmentation at the original resolution.

TransUNet. The TransUNet (Chen et al., 2021) architecture consists of a hybrid CNN-ViT encoder (Fig. 3) that was designed to overcome the limitations of traditional CNN-based models which may struggle with capturing global context information, by incorporating the powerful self-attention mechanism involved in ViT models (Dosovitskiy et al., 2020). The 3D CNN feature extractor captures the local features that are later fed into the Transformer-based layers to capture global feature representation by applying several multi-head self-attention (MHSA) blocks. The self-attention mechanism used in Transformers allows the network to attend to relevant features across the entire image, thus capturing long-range dependencies and global context information. By combining this with the local feature extraction of the CNN, TransUNet is able to effectively capture both local and global context information and to produce highly accurate segmentation masks. Finally, the output of the encoder network is then passed to a decoder consisting of a series of upsampling and convolutional layers that progressively increase the spatial resolution of the features and produce the final delineation masks.

MedT. The MedT architecture is constructed using a gated axial Transformer layer, which effectively addresses the computational complexity associated with computing the self-attention mechanism (Valanarasu et al., 2021). Furthermore, the incorporation of a gating mechanism enables precise control over the impact of the learned relative positional encodings on encoding non-local context. If a relative positional encoding is accurately learned, the gating mechanism assigns it a higher weight. MedT achieves image segmentation through the use of two branches (Fig. 4): a global branch that operates on the original resolution of the image, and a local branch that operates on patches of the image. In the local branch, the image is divided into 16 patches of size $H/4 \times W/4$ of the original image. Each patch is then processed through the network, and the output feature maps are re-sampled based on their location to obtain the output feature maps. The output feature maps from both branches are combined and passed through a 1×1 convolution layer to produce the final segmentation mask. This

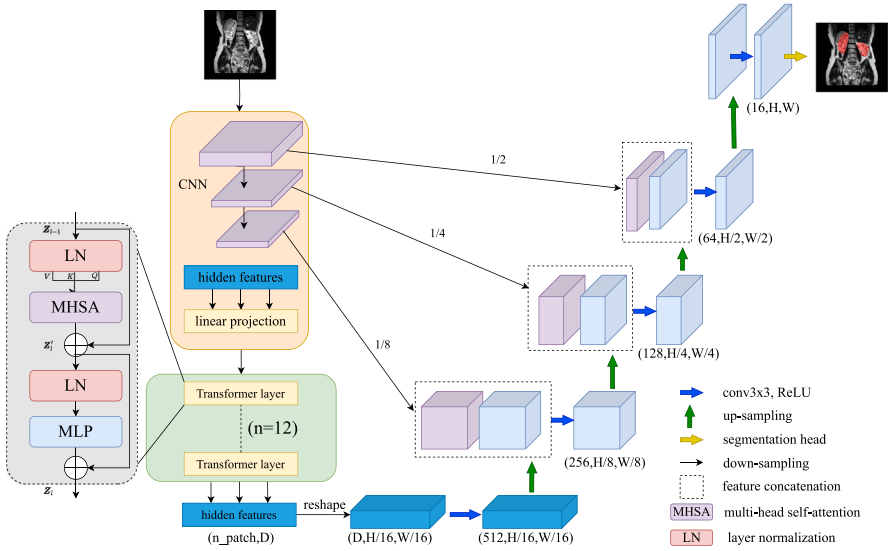

Fig. 3. TransUNet (Chen et al., 2021), hybrid CNN-Transformer encoder–decoder architecture.

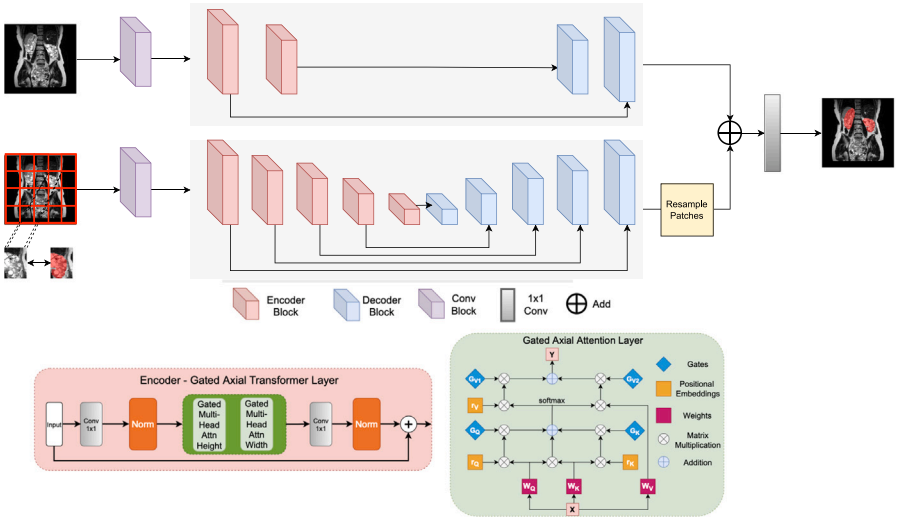

Fig. 4. MedT (Valanarasu et al., 2021), hybrid CNN-Transformer encoder–decoder architecture.

approach enhances performance as the global branch focuses on high-level information while the local branch concentrates on capturing finer details.

SwinUNetV2. SwinUNetV2 is a UNet-like shape architecture with a SwinV2 Transformer-based encoder and a CNN-based decoder (Liu et al., 2022). Unlike ViT Transformer, Swin Transformers introduce a sliding window strategy to only compute the attention within local windows, thus reducing the computational cost of global multi-head self-attention. Swin Transformers gradually decrease the initial number of tokens by implementing patch merging layers as the network gets deeper. This attribute delivers a hierarchical representation similar to the way CNN behaves. In practice, Swin Transformer consists of four stages and performs 2×2 spatial downsampling in the patch merging layer of each stage (Fig. 5). The merging layer concatenates the features of each group of 2×2 neighboring patches and applies a linear transformation to reduce the number of concatenated features to half their dimension. The SwinV2 Transformer proposes three main modifications to better scale up model capacity and window resolution. First, a res-post-norm configuration is suggested to replace the previous pre-norm configuration. Second, a scaled cosine attention mechanism is proposed to replace the original dot product attention. Finally, a log-spaced continuous relative position bias approach is introduced to replace the previous parameterized approach. These adaptations make it easier for the model to scale up capacity and improve its transferability across different resolutions. Specifically, the res-post-norm configuration and the scaled cosine attention mechanism enhance the model's ability to handle larger input sizes and more complex tasks, while the log-spaced continuous relative position bias approach ensures that the model can effectively capture long-range dependencies across spatial locations.

Segmenter. The Segmenter is an encoder–decoder architecture based on a Transformer (Fig. 6) that is designed to map a sequence of patch embeddings to pixel-level class annotations (Strudel et al., 2021). The encoder follows the design proposed in the original ViT approach (Dosovitskiy et al., 2020). In this design, the input is reshaped into a sequence of flattened, uniform, and non-overlapping patches. The number of patches obtained becomes the effective input length of the Transformer. To project the patches into a d-dimensional embedding space, a linear layer is used, and a 1D learnable patch position embedding is added to retain positional information. The projected embeddings are then passed into the Transformer-based layers, which

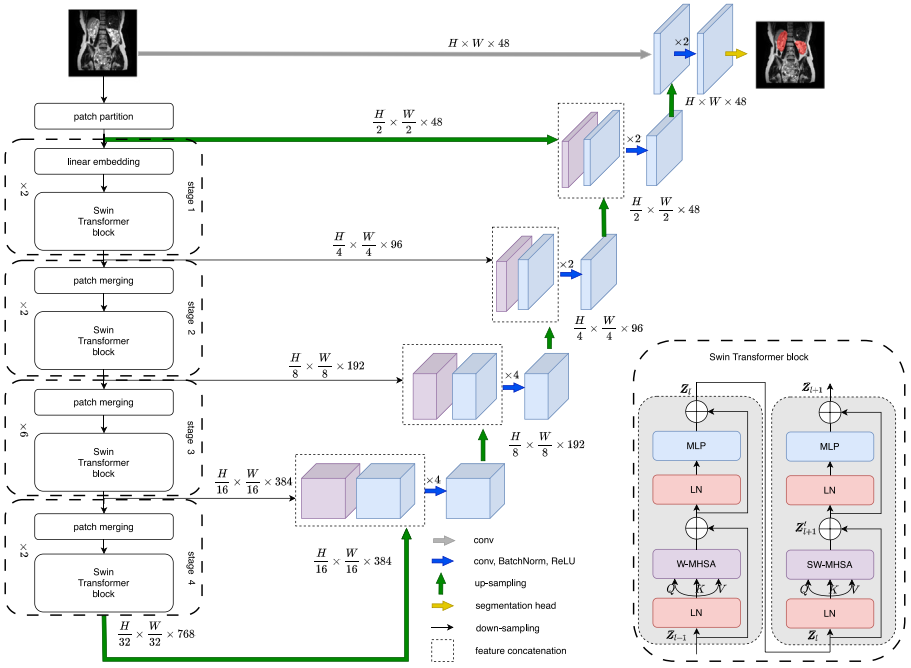

Fig. 5. SwinUNetV2 (Liu et al., 2022), hybrid CNN-Transformer encoder–decoder architecture.

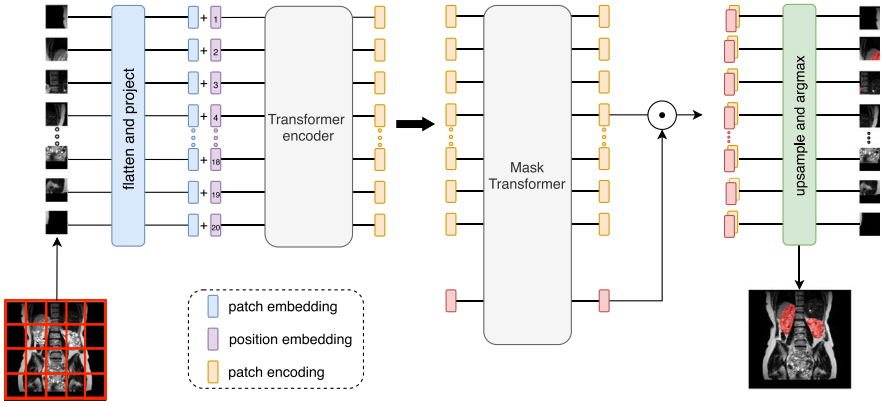

Fig. 6. Segmenter (Strudel et al., 2021), Transformer-based encoder–decoder architecture.

include h multi-head attention and multi-layer perceptron (MLP) sub-layers. The resulting patch embeddings are combined with an additional learnable class embedding in the Transformer decoder. The decoder generates the final mask by calculating the scalar product between the L2-normalized final output patches and the learnable class embedding. Each patch then represents the probability of belonging to a specific class. The resulting mask sequences are then reshaped and linearly upsampled to the original image size. To get the pixel-wise class scores forming the final segmentation, a softmax is applied on the class dimension, followed by a normalization layer.

3. Experiments

3.1. Imaging dataset

Data were collected from the Genkyst[1] study, a regional cohort involving nephrologists working in private and public nephrology centers in the West of France. It registered imaging, clinical and genetic data of all consenting patients with ADPKD from this area. An institutional review board approval was obtained for this study. Informed consents were obtained for all subjects.

In practice, a set of 118 MR images arising from 112 patients with ADPKD was considered. The MR images were coronal single shot fast spin echo T2 sequences, acquired in various centers and with various devices. Among the 118 T2 images, 34 were identified as atypical because of: 1 - artifacts (e.g. noise), 2 - unknown genetic mutation different from *PKD1t*, *PKD1nt*, or *PKD2*, or 3 - unknown genetic mutation (approximately 5% of genetically unresolved patients). These images were therefore used for training purposes only. Among the 84 remaining images, 10 images were identified from patients for which 2 examinations were available. Regarding genetic class distribution (Fig. 7*a*), 42% of patients had a *PKD1nt* mutation, 52% for *PKD1t*, 9% for *PKD2* and 15% for other or unknown genetic mutations. The Mayo class distribution, display in Fig. 7*b*, was as follows: 6% of patients for 1A, 21% for 1B, 54% for 1C, 22% for 1D, 10% for 1E, 2% for 2. Such information was missing for 3% of patients. The average age of patient was 47.1 years old in average, with a standard deviation of 14.2. 41.5% (resp. 58.5%) of patients were males (resp. females). The employed dataset covered a large range of various kidney volumetries (Fig. 8), from around 200 to > 2000 mL per kidney.

[1] https://clinicaltrials.gov/ct2/show/NCT02887729

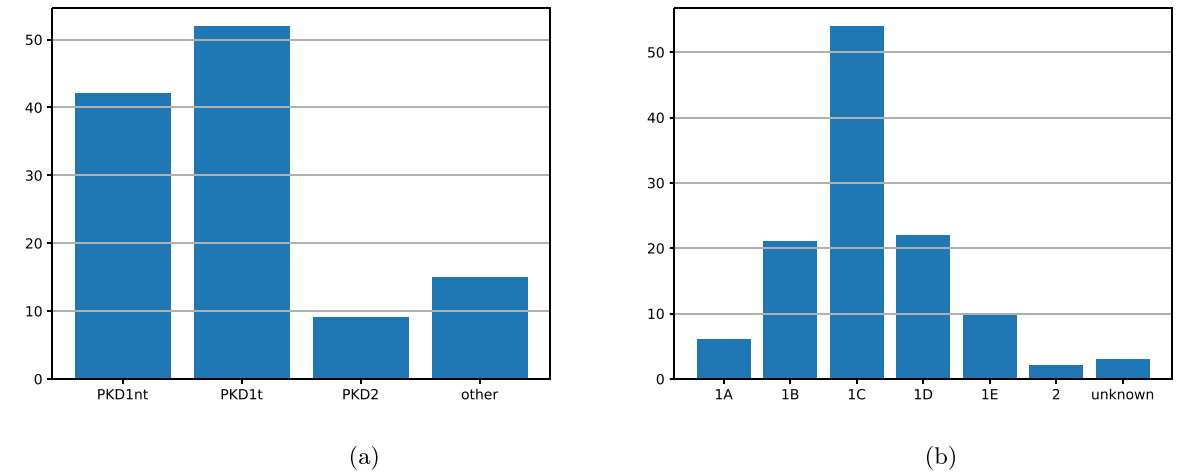

Fig. 7. Dataset visualization. Class distribution with (a) non-truncating (*PKD1nt*), truncating (*PKD1t*) *PKD1*, *PKD2* and other mutations, (b) 1A, 1B, 1C, 1D, 1E and 2 Mayo classes.

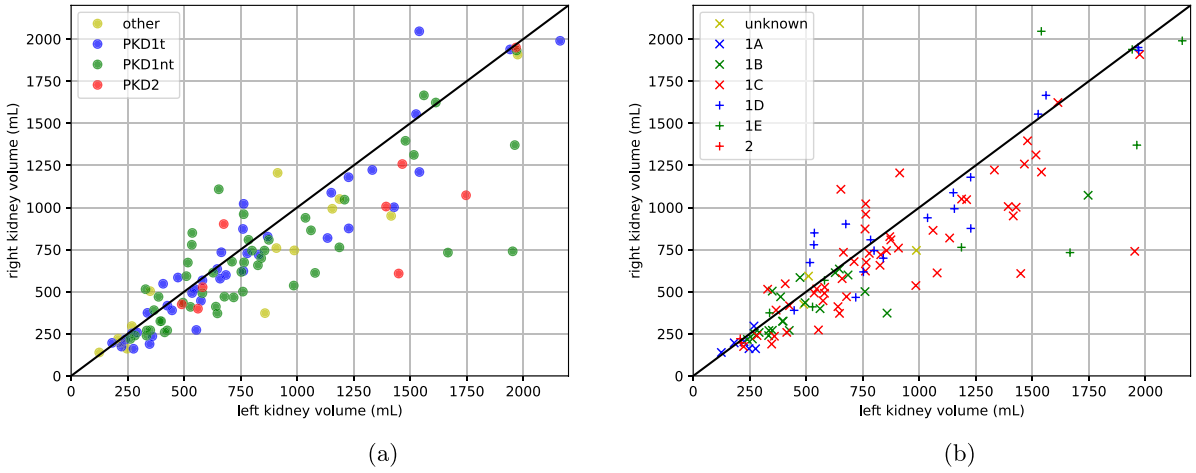

Fig. 8. Dataset visualization. Left kidney volume with respect to right kidney volume across the whole dataset with (a) truncating and non-truncating *PKD1* (respectively *PKD1t* and *PKD1nt*), *PKD2* and other gene mutations, (b) 1A, 1B, 1C, 1D, 1E and 2 Mayo classes.

The images had a reconstructed matrix size of 512 × Y × 512, with Y large enough to cover the full extent of both kidneys. Image voxel sizes were most commonly of the order of 0.8 mm coronal in-plane with a slice thickness of approximately 4 mm. For training and evaluation purposes, left and right kidneys were manually outlined by clinical experts using the ITKSnap[2] software. In practice, two clinical research associates (from the clinical research unit attached to the nephrology department of University Hospital of Brest, France) performed the manual kidney delineation task. Subsequently, a senior nephrologist (with 15 years of experience) carefully reviewed and validated these annotations to verify their precision and quality.

3.2. Implementation details

The different networks (v19pUNet, TransUNet, MedT, SwinUNetV2, Segmenter) and learning schemes (BO, IND, DT) were implemented in 2D (i.e. by considering coronal slices) using Pytorch. Experiments were performed using a single NVIDIA 1080Ti GPU, with 11 Gb of RAM. Deep networks were trained with data augmentation (i.e. random affine transforms), 200 epochs, an Adam optimizer, a fuzzy Dice score as loss function, a learning rate of 10^{-3} for MedT and 10^{-5} for other architectures. For BO and IND learning schemes, the batch size was set to 16, 16, 5, 16 and 4 images for v19pUNet, TransUNet, MedT, SwinUNetV2 and Segmenter respectively. 8, 8, 4, 8 and 3 were respectively used for DT to avoid memory issues. 2D coronal slices were reshaped to 256 × 256.

3.3. Segmentation assessment

Each configuration (i.e, a given network architecture in one given learning scheme) was performed in a 5-fold cross-validation manner. Each fold employed 88 (or 89 depending on the fold) examinations for training, 12 for validation and 17 (or 18) for test, hence performing a 75%, 10% and 15% split. The splitting process between training, validation, and test sets was guided by the patient independence criterion such that two examinations from the same patient were assigned to only one of the subsets.

To evaluate the performance of the different deep models, we compared ground truth GT and prediction P masks, respectively defined by the surface S_{GT} and S_P, through Dice similarity coefficient (DSC defined as $\frac{2|GT \cap P|}{|GT|+|P|}$), absolute volume difference (AVD expressed as $\frac{||GT|-|P||}{|GT|}$) as well as average symmetric surface (ASSD = $\frac{1}{|S_{GT}|+|S_P|}(\sum_{s \in S_{GT}} d(s, S_P) + \sum_{s \in S_P} d(s, S_{GT}))$ where $d(s, S_k) = \min_{s_k \in S_k} \|s - s_k\|$) and Haussdorf (HD = $\max(h(GT, P), h(P, GT))$ with $h(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\|$) distances. Final metric scores were averaged among the 5 folds to provide a reliable performance trend.

[2] http://www.itksnap.org/

Table 1
Quantitative assessment of automatic polycystic kidney segmentation. Comparison between v19pUNet (Conze et al., 2020), TransUNet (Chen et al., 2021), MedT (Valanarasu et al., 2021), Segmenter (Strudel et al., 2021) and SwinUNetV2 (Liu et al., 2022) architectures for BO, IND and DT strategies. Evaluation is provided for both kidneys (BK) using DSC, AVD, ASSD and HD metrics. Bold results indicate the best scores among all architectures and learning schemes while underline results indicate the best results for a given network (among BO, IND and DT strategies).

		BK			
		DSC	AVD	ASSD	HD
v19pUNet (Conze et al., 2020)	BO	92.2 ± 3.62	0.12 ± 0.10	1.90 ± 1.46	34.3 ± 18.8
	IND	92.4 ± 3.85	0.11 ± 0.10	1.76 ± 1.74	32.9 ± 20.7
	DT	93.2 ± 2.78	0.09 ± 0.06	1.44 ± 1.13	28.6 ± 18.1
TransUNet (Chen et al., 2021)	BO	92.9 ± 2.92	0.09 ± 0.06	1.50 ± 1.34	28.5 ± 18.6
	IND	92.8 ± 3.07	0.10 ± 0.06	1.54 ± 1.36	29.0 ± 20.2
	DT	93.2 ± 2.75	0.09 ± 0.05	1.42 ± 1.24	27.8 ± 18.3
MedT (Valanarasu et al., 2021)	BO	91.8 ± 4.26	0.09 ± 0.06	1.86 ± 1.62	31.2 ± 18.8
	IND	91.6 ± 4.12	**0.08** ± 0.04	1.96 ± 1.51	33.6 ± 20.0
	DT	92.4 ± 3.94	**0.08** ± 0.04	1.66 ± 1.41	29.4 ± 17.1
Segmenter (Strudel et al., 2021)	BO	92.0 ± 3.86	0.08 ± 0.06	1.82 ± 1.64	31.2 ± 19.2
	IND	91.6 ± 4.80	0.09 ± 0.06	1.89 ± 1.73	31.8 ± 19.5
	DT	92.1 ± 4.22	0.09 ± 0.06	1.84 ± 1.74	32.5 ± 21.4
SwinUNetV2 (Liu et al., 2022)	BO	93.1 ± 2.87	0.10 ± 0.06	1.55 ± 1.38	32.2 ± 21.3
	IND	92.9 ± 3.14	0.10 ± 0.07	1.58 ± 1.37	31.0 ± 18.9
	DT	**93.4** ± 2.76	0.09 ± 0.06	**1.35** ± 1.22	**26.8** ± 17.2

Table 2
Statistical analysis in DSC for both kidneys (BK) between DT and BO/IND strategies using v19pUNet (Conze et al., 2020), TransUNet (Chen et al., 2021), MedT (Valanarasu et al., 2021), Segmenter (Strudel et al., 2021) and SwinUNetV2 (Liu et al., 2022) through Student's paired t-tests. Bold p-values (<0.05) highlight statistically significant results.

	v19pUNet (Conze et al., 2020)	TransUNet (Chen et al., 2021)	MedT (Valanarasu et al., 2021)	Segmenter (Strudel et al., 2021)	SwinUNetV2 (Liu et al., 2022)
	DT				
BO	$\mathbf{1.19 \times 10^{-7}}$	$\mathbf{3.47 \times 10^{-5}}$	$\mathbf{3.02 \times 10^{-4}}$	3.07×10^{-1}	$\mathbf{2.19 \times 10^{-2}}$
IND	$\mathbf{1.78 \times 10^{-6}}$	$\mathbf{1.44 \times 10^{-7}}$	$\mathbf{9.31 \times 10^{-8}}$	$\mathbf{4.44 \times 10^{-5}}$	$\mathbf{5.96 \times 10^{-4}}$

4. Results and discussion

Results are reported and discussed by taking into consideration the segmentation performance for both kidneys (BK) as well as for individual left (LK) and right (RK) kidneys through DSC, AVD, ASSD and HD metrics (Section 3.3). In particular, the assessment focuses on both standard (BO, IND) versus dual-task (DT) learning schemes (Section 4.1), while also comparing the performance of convolutional (v19pUNet) versus hybrid CNN/Transformer-based (TransUNet, MedT, SwinUNetV2) and Transformer-based (Segmenter) architectures (Section 4.2). The performance achieved for each architecture/learning scheme couple at the level of both kidneys (BK) is summarized in Table 1 whereas Table 3 provides a more detailed analysis by studying the delineation accuracy for each kidney (LK, RK). This analysis is supplemented by statistical analyses (Table 2 and 4) through Student paired t-tests with bilateral distribution, Bland-Altman and concordance curves (Fig. 9, 10 and 12) as well as qualitative results (Fig. 11, 13) to provide a thorough evaluation of the models' performance and facilitates informed decision-making dealing with their potential integration into clinical routine.

4.1. Standard versus dual-task learning schemes

When studying the delineation performance for both kidneys simultaneously (BK), it appears that the models and their respective learning scheme range from an average DSC of 91.6% to 93.4%. The worst performance is observed when using MedT and Segmenter architectures trained through the IND strategy, whereas the best performance is achieved with the SwinUNetV2 model employing the DT configuration. Notably, whatever the used network, the employment of the multi-task learning scheme (i.e. DT) consistently yields the highest overall DSC performance with respect to BO and IND strategies. This observation remains consistent across the other metrics (AVD, ASSD and HD) with the exception of the Segmenter architecture where AVD, ASSD and HD metric values are slightly better in BO configuration (e.g. 1.82 mm versus 1.84 mm in ASSD). Except when comparing DT and BO strategies for Segmenter, above conclusions (DT > BO and DT > IND) are further supported by the statistical analysis provided in Table 2 through Student's paired t-tests.

Two main findings arise from Table 1. First, task-specific decoders are better able to process features from a single joint encoder $f_{LK+RK}(\cdot)$ (DT) than from kidney-specific encoders (IND), i.e. $f_{LK}(\cdot)$ and $f_{RK}(\cdot)$. Encoding features simultaneously extracted from both kidneys hence improves the per-kidney delineation tasks, making DSC scores increasing from 92.4% (resp. 91.6%) to 93.2% (resp. 92.4%) and ASSD values decreasing from 1.76 mm (resp. 1.96 mm) to 1.44 mm (resp. 1.66 mm) for v19pUNet (resp. MedT). Second, exploiting decoding branches, i.e. $g_{LK}(\cdot)$ and $g_{RK}(\cdot)$, respectively targeting left and right kidneys (DT), provides better results than employing a single joint decoder $g_{BK}(\cdot)$ (BO). Although more computationally costly than BO, the DT scheme allows the reconstruction task to be further specialized. Thus, the DSC (resp. HD) metric is increased (resp. reduced) from 92.2% (resp. 34.3 mm) to 93.2% (resp. 28.6 mm) for v19pUNet and from 93.1% (resp. 32.2 mm) to 93.4% (resp. 26.8 mm) for SwinUNetV2. It is noteworthy to mention that the BO strategy reaches a slight improvement over the IND learning scheme for all models, except for v19pUNet where the IND demonstrates superior performance. Finally, models that include Transformer blocks (e.g. TransUNet, MedT, Segmenter, SwinUNetV2) demonstrate a closer performance across learning schemes than v19pUNet, particularly when the DSC metric is considered.

The same conclusions arise when studying the accuracy of contours reached for each kidney (Table 3). Note that for the BO learning scheme, kidney-specific results were obtained by identifying the two largest connected components and classifying them (left or right kidney) with respect to their minimal position in X-axis. When focusing exclusively on the left kidney (LK), the metric values demonstrate a range spanning from 92.6% to 94.1% for DSC, 0.10% to 0.07% for AVD, 1.70 mm to 1.14 mm for ASSD, and 26.6 mm to 20.2 mm for HD. The DT strategy consistently delivers the best performance across all models and metrics. However, the magnitude of the difference varies

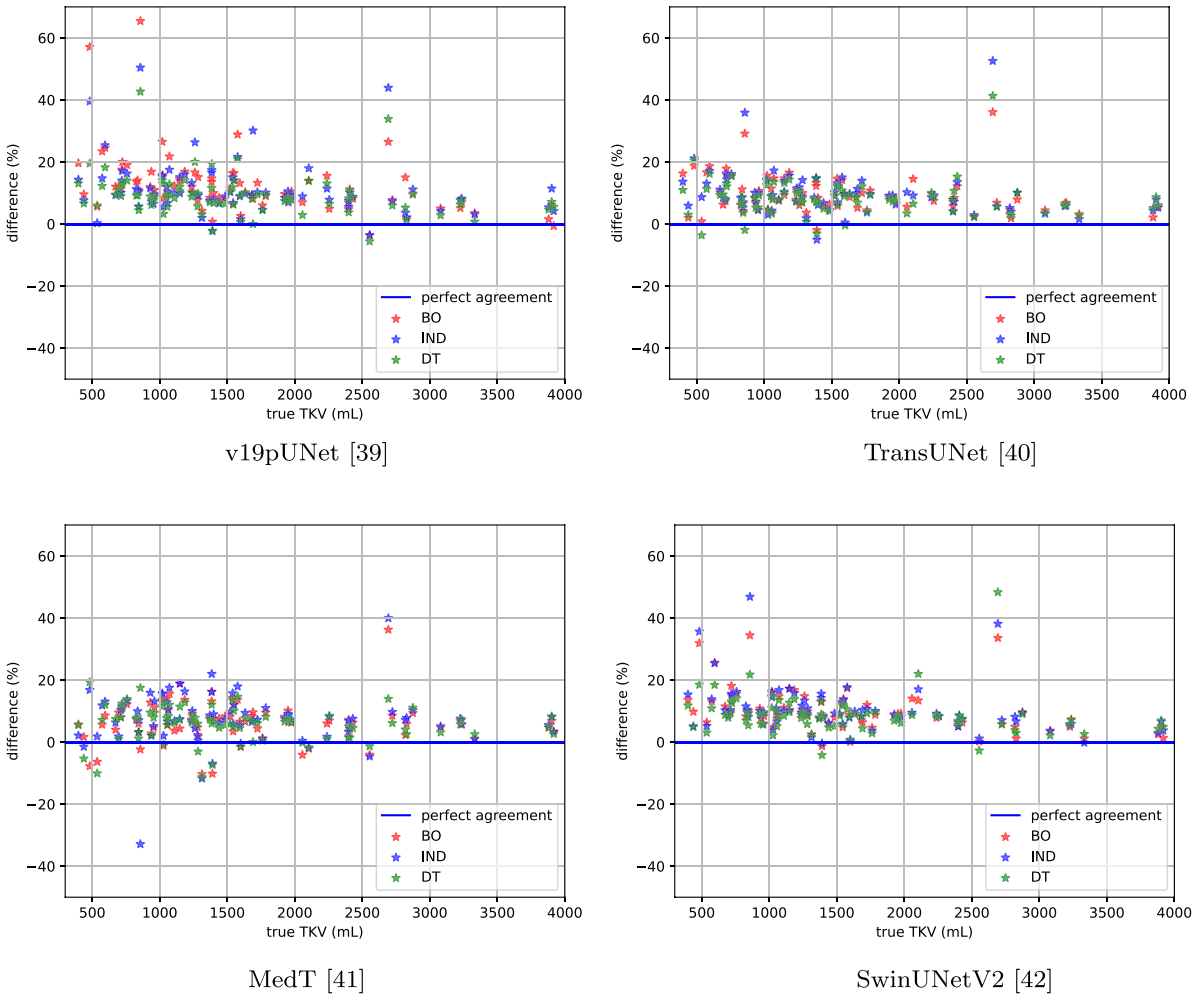

Fig. 9. Bland-Altman analysis of the percent difference of TKV measurements obtained by v19pUNet (Conze et al., 2020), TransUNet (Chen et al., 2021), MedT (Valanarasu et al., 2021) and SwinUNetV2 (Liu et al., 2022) architectures for BO, IND and DT strategies with respect to reference TKV estimated from ground truth annotations.

Table 3
Quantitative assessment of automatic polycystic kidney segmentation. Comparison between v19pUNet (Conze et al., 2020), TransUNet (Chen et al., 2021), MedT (Valanarasu et al., 2021), Segmenter (Strudel et al., 2021) and SwinUNetV2 (Liu et al., 2022) architectures for BO, IND and DT strategies. Evaluation is provided for left (LK) and right (RK) kidneys using DSC, AVD, ASSD and HD metrics. Bold results indicate the best scores among all architectures and learning schemes while underline results indicate the best results for a given network (among BO, IND and DT strategies).

		LK				RK			
		DSC	AVD	ASSD	HD	DSC	AVD	ASSD	HD
v19pUNet (Conze et al., 2020)	BO	93.4 ± 3.01	0.10 ± 0.08	1.36 ± 0.76	22.7 ± 10.5	91.2 ± 5.88	0.14 ± 0.13	2.02 ± 2.03	28.6 ± 17.9
	IND	92.8 ± 4.23	0.11 ± 0.13	1.70 ± 2.11	26.6 ± 19.4	91.7 ± 5.41	0.13 ± 0.13	1.79 ± 1.77	27.7 ± 18.7
	DT	93.9 ± 2.50	0.08 ± 0.06	1.24 ± 1.02	22.4 ± 13.6	92.2 ± 4.91	0.12 ± 0.13	1.61 ± 1.83	24.6 ± 16.7
TransUNet (Chen et al., 2021)	BO	93.6 ± 2.09	0.08 ± 0.05	1.21 ± 0.52	20.5 ± 9.48	91.8 ± 5.57	0.12 ± 0.16	1.78 ± 2.34	26.0 ± 20.5
	IND	93.6 ± 2.29	0.08 ± 0.05	1.22 ± 0.55	20.9 ± 10.6	91.6 ± 5.90	0.13 ± 0.20	1.82 ± 2.29	26.1 ± 20.7
	DT	93.9 ± 1.89	**0.07** ± 0.04	1.16 ± 0.48	**20.2** ± 10.1	92.2 ± 5.36	0.12 ± 0.17	1.61 ± 2.16	24.0 ± 18.5
MedT (Valanarasu et al., 2021)	BO	92.6 ± 3.44	0.08 ± 0.06	1.55 ± 1.00	23.8 ± 11.8	90.7 ± 6.86	0.12 ± 0.18	2.08 ± 2.58	27.0 ± 20.1
	IND	92.6 ± 3.77	0.07 ± 0.05	1.56 ± 1.02	24.0 ± 11.0	90.3 ± 6.49	0.12 ± 0.11	2.32 ± 2.52	29.8 ± 21.4
	DT	93.1 ± 3.33	**0.07** ± 0.04	1.42 ± 1.08	22.7 ± 11.5	91.4 ± 5.92	**0.10** ± 0.11	1.87 ± 2.09	26.4 ± 19.3
Segmenter (Strudel et al., 2021)	BO	92.7 ± 2.83	0.08 ± 0.05	1.55 ± 0.97	23.9 ± 11.3	91.0 ± 6.48	0.11 ± 0.15	2.04 ± 2.69	26.8 ± 20.4
	IND	92.7 ± 3.13	0.07 ± 0.05	1.49 ± 0.91	23.0 ± 10.6	90.1 ± 8.74	0.14 ± 0.19	2.26 ± 3.00	28.2 ± 21.5
	DT	93.0 ± 2.62	0.07 ± 0.05	1.43 ± 0.80	22.5 ± 11.5	90.9 ± 7.15	0.13 ± 0.18	2.18 ± 2.89	28.7 ± 22.8
SwinUNetV2 (Liu et al., 2022)	BO	93.9 ± 2.20	0.08 ± 0.05	1.21 ± 0.75	23.2 ± 12.3	92.0 ± 5.32	0.13 ± 0.14	1.86 ± 2.38	28.1 ± 22.4
	IND	93.8 ± 2.28	0.08 ± 0.05	1.21 ± 0.70	22.3 ± 11.5	91.6 ± 5.77	0.14 ± 0.17	1.92 ± 2.37	27.8 ± 21.1
	DT	**94.1** ± 2.16	**0.07** ± 0.04	**1.14** ± 0.61	20.7 ± 10.8	**92.4** ± 5.48	0.12 ± 0.17	**1.54** ± 2.18	**23.0** ± 17.2

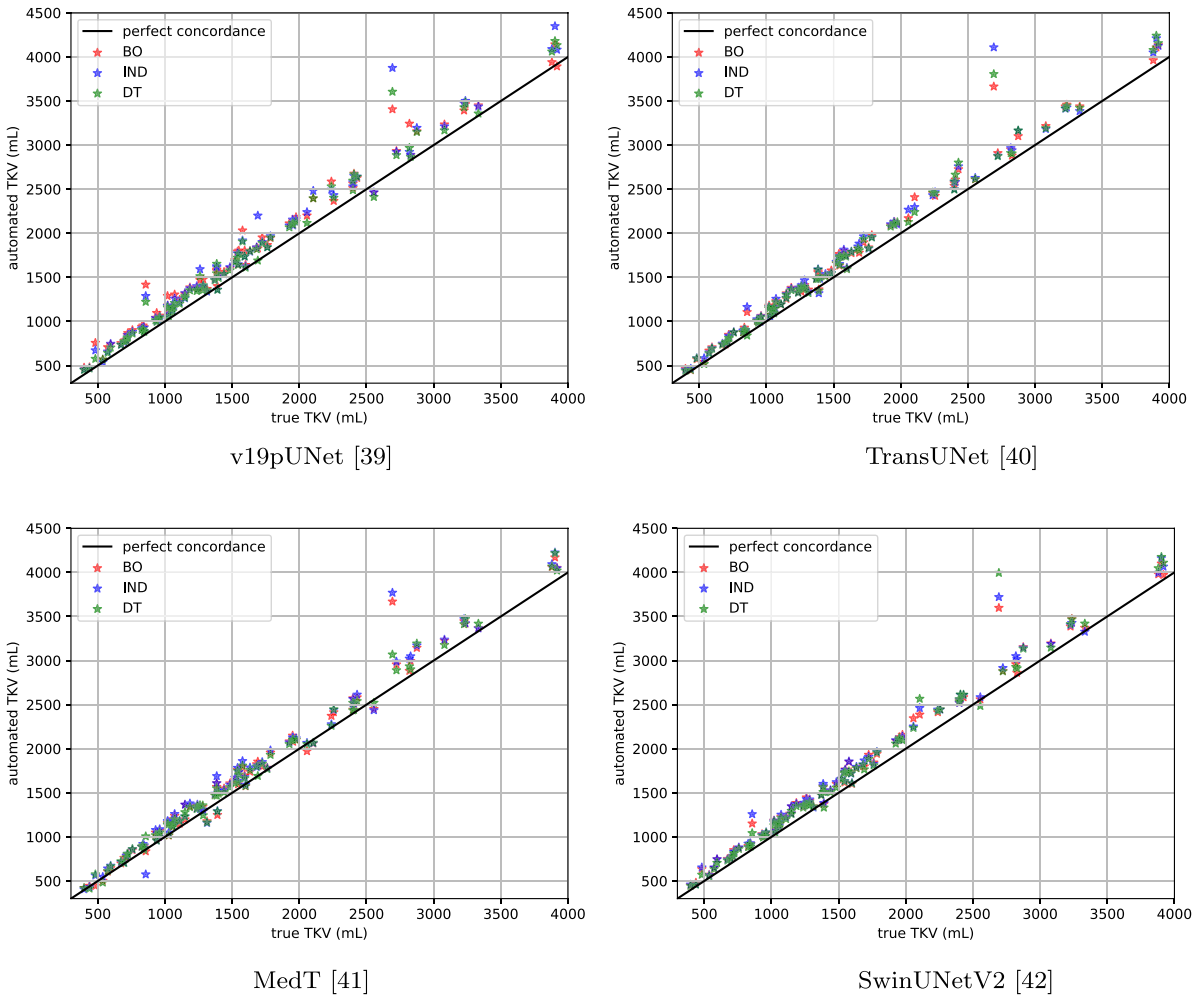

Fig. 10. Concordance analysis between TKV measurements obtained by v19pUNet (Conze et al., 2020), TransUNet (Chen et al., 2021), MedT (Valanarasu et al., 2021) and SwinUNetV2 (Liu et al., 2022) architectures for BO, IND and DT strategies and reference TKV estimated from ground truth annotations.

depending on the model. For instance, the v19pUNet model exhibits a larger disparity between the standard (BO, IND) and multi-task (DT) strategies up to 1.1% gap in DSC, 0.03% in AVD, 0.26 mm in ASSD, and 4.2 mm in HD for LK. On the contrary, metrics values for right kidney (RK) span from 90.1% to 92.4% for DSC, 0.14% to 0.10% for AVD, 2.32 mm to 1.54 mm for ASSD and 29.8 mm to 23.0 mm for HD. These findings provide evidence that right kidney segmentation presents a greater challenge compared to left kidney delineation, as manifested by an approximate 2% decrease in DSC. This underscores the heightened difficulty in accurately delineating the right kidney, especially due to its vicinity with the liver which often present cysts in patients with ADPKD. Similar to the observations for left kidney (LK) and apart from TransUNet in AVD and Segmenter in all metrics, the involved models consistently achieve their highest performance when employing the DT learning scheme. However, the magnitude of this difference varies depending on the employed deep architecture.

The concordance between predicted and ground truth TKV estimates demonstrates a stronger correlation for DT than for BO and IND with individual estimations closer to the lines of perfect agreement (Fig. 9) and concordance (Fig. 10). A tendency to over-estimate kidney volumes is revealed across all tested deep networks and learning schemes (Fig. 9, 10). Visually, the contours obtained through the DT learning scheme more closely follow the ground truth delineations compared to both BO and IND strategies (Fig. 11). A better distinction between hepatic and renal cysts is also noteworthy, which further supports the benefits of task-specific decoders processing features arising from a joint encoder.

4.2. Convolutional versus transformers-based architectures

Several notable insights emerge when comparing the performance of CNN versus Transformer-based architectures for polycystic kidney delineation (Table 1). Particularly, among the examined models, hierarchical Transformers (i.e. SwinUNetV2) showcase superior performance across the different learning schemes. However, it is essential to highlight that the hybrid CNN-Transformer model, TransUNet, and the pure convolutional network, v19pUNet, closely approach the performance of SwinUNetV2, particularly in the DT scenario, with differences in DSC scores falling within a margin of 0.2%. Unlike other network pairs, DSC comparisons for both kidneys (BK) between v19pUNet, TransUNet and SwinUNetV2 does not reach statistical significance (Table 4). This difference is much more pronounced in both BO and IND configurations (e.g. 92.2% for v19pUNet against 93.1% for SwinUNetV2 in DSC using the BO scheme). In regard to the remaining metrics, SwinUNetV2 exhibits a decrement of at least 1.0 mm in HD compared to its competitors. Conversely, the CNN-based models with attention mechanisms, MedT, and the pure Transformer-based network, Segmenter, display the poorest performance in the DT scenario. Notably, both networks exhibit a decrease of 1.0% and 1.3% in DSC respectively, when compared to the best-performing model. Contrary to other models, the MedT network tends to provide downward TKV predictions according to both agreement and concordance analyses (Fig. 12). It reaches the best AVD score with 0.08% in BK. The Segmenter architecture reveals strong limitations in accurately capturing fine details and fully delineating the

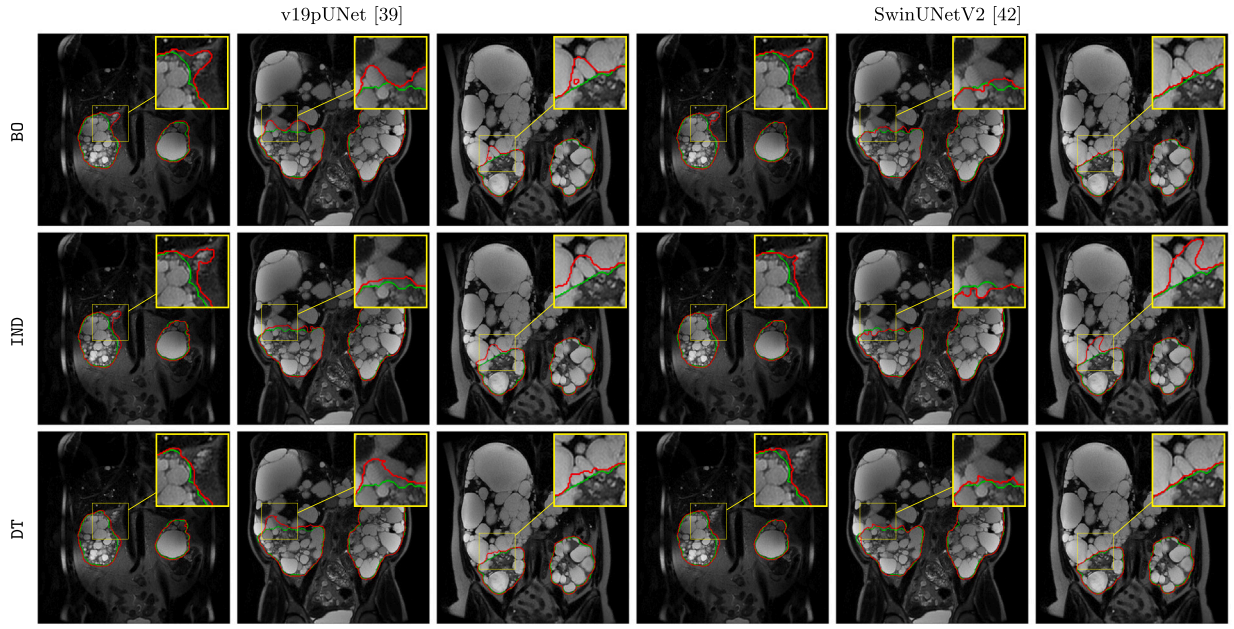

Fig. 11. Visual results using v19pUNet (Conze et al., 2020) and SwinUNetV2 (Liu et al., 2022) trained with BO, IND and DT learning strategies. Ground truth and estimated contours are resp. in green and red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Statistical analysis in DSC for both kidneys (BK) between v19pUNet (Conze et al., 2020), TransUNet (Chen et al., 2021), MedT (Valanarasu et al., 2021), Segmenter (Strudel et al., 2021) and SwinUNetV2 (Liu et al., 2022) with DT strategy through Student's paired t-tests. Bold p-values (<0.05) highlight statistically significant results.

	v19pUNet (Conze et al., 2020)	TransUNet (Chen et al., 2021)	MedT (Valanarasu et al., 2021)	Segmenter (Strudel et al., 2021)	SwinUNetV2 (Liu et al., 2022)
v19pUNet (Conze et al., 2020)	.	9.96×10^{-1}	$\mathbf{3.96 \times 10^{-5}}$	$\mathbf{3.63 \times 10^{-6}}$	1.59×10^{-1}
TransUNet (Chen et al., 2021)	–	.	$\mathbf{5.73 \times 10^{-4}}$	$\mathbf{1.26 \times 10^{-5}}$	1.02×10^{-1}
MedT (Valanarasu et al., 2021)	–	–	.	$\mathbf{1.46 \times 10^{-2}}$	$\mathbf{1.55 \times 10^{-5}}$
Segmenter (Strudel et al., 2021)	–	–	–	.	$\mathbf{8.02 \times 10^{-7}}$
SwinUNetV2 (Liu et al., 2022)	–	–	–	–	.

target structures, as evidenced by the high values observed for both ASSD and HD metrics (Table 1, Table 3) as well as the visual results depicted in Fig. 13. This issue aligns with previous findings in the literature regarding Transformers-based networks for segmentation (Liu et al., 2021) which utilize patch-based approaches for a task that requires dense prediction at the pixel level. The use of patches can result in the omission of certain regions or the inclusion of non-target regions as part of the segmented structure, leading to sub-optimal segmentation robustness.

These findings remain consistent with the ones revealed by the per-kidney metric values provided in Table 3. Among the models evaluated, SwinUNetV2 consistently outperforms the other models, showcasing its superiority. However, it is worth noting that MedT performs similarly or better according to the AVD metric, with scores of 0.07 and 0.10 for LK and RK, respectively. A more detailed analysis of Fig. 13 allows for better observations. Notably, the Segmenter model, a full Transformer network, struggles with correct segmentation, leading to under-segmentation issues (1B, 1C, 1D, 1E). Although Transformers excel at capturing long-range dependencies, they fall short when it comes to preserving fine details. On the other hand, MedT, which combines a global and local branch, encounters challenges in capturing intricate details and experiences problems related to adjacent liver cysts (1B) and over-segmentation (1D). While v19pUNet shows improved performance in capturing fine details, it faces difficulties in accurately delineating the concavity of the left and right kidneys (1B, 1C, 1D). The hybrid TransUNet model, with its fusion of hierarchical features (CNN) and long-range modeling (Transformers), demonstrates a better representation of kidney concavities but still displays minor contouring errors. It is interesting to note that SwinUNetV2 is the only model able to accurately capturing the cyst belonging to the upper part of the left kidney in the 1E sample from Fig. 13. More globally, it emerges as the front-runner, boasting the highest accuracy in kidney delineation (Table 3), providing TKV estimates that are consistent with ground truth (Fig. 12) and excelling in capturing the kidney's shape (Fig. 13).

Our study underscores the substantial advancements made by hierarchical Transformers, as they exhibit a strong proficiency for the kidney MR delineation task. Their ability to capture multi-scale information and leverage hierarchical representations significantly contributes to their superior performance. Additionally, the hybrid TransUNet model and the pure convolutional network v19pUNet also demonstrate competitive performance, indicating the continued relevance and effectiveness of CNN-based approaches.

5. Conclusion

In this work, we successfully addressed fully-automated kidney delineation for patients with autosomal-dominant polycystic kidney disease. In particular, we contributed to the advancement of polycystic kidney segmentation from MR scans by investigating and comparing

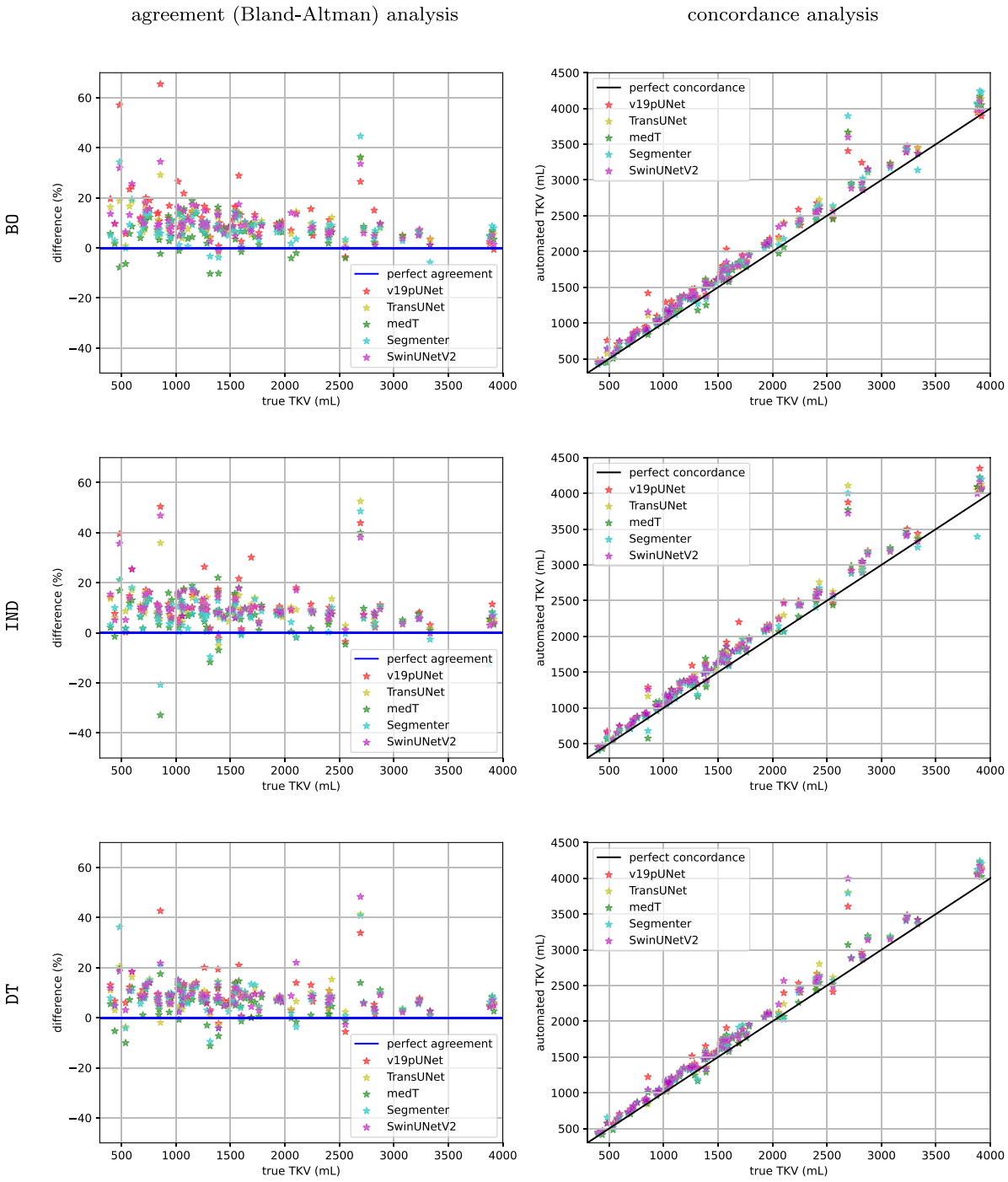

Fig. 12. Bland-Altman and concordance analyses between TKV measurements obtained by BO, IND and DT strategies using v19pUNet (Conze et al., 2020), TransUNet (Chen et al., 2021), MedT (Valanarasu et al., 2021), Segmenter (Strudel et al., 2021) and SwinUNetV2 (Liu et al., 2022) architectures and reference TKV estimated from ground truth annotations.

different network architectures including CNN-based, Transformers-based, and hybrid CNN/Transformers-based models. Furthermore, we emphasized the importance of multi-task learning in medical image analysis by proposing a dual-task learning scheme, where a common feature extractor is followed by per-kidney decoders. To evaluate and compare the performance of different network architectures and learning strategies, we conducted comprehensive experiments on a heterogeneous MR imaging dataset collected from 112 patients. The results of our study provided valuable insights into the effectiveness of hybrid CNN/Transformers trained in a dual-task fashion. Moving forward, future research will focus on addressing the challenges posed by the heterogeneous distributions of cysts throughout the abdomen and improving the robustness of deep models to differentiate polycystic kidneys from surrounding cyst-containing organs. Additionally, the use of longitudinal data could further enhance the predictive capabilities of hybrid CNN/Transformer-based segmentation tools. Conducting a multi-center study involving multiple institutions and leveraging a service-oriented network-based infrastructure to bolster service continuity should also deserve further investigation in the near future.

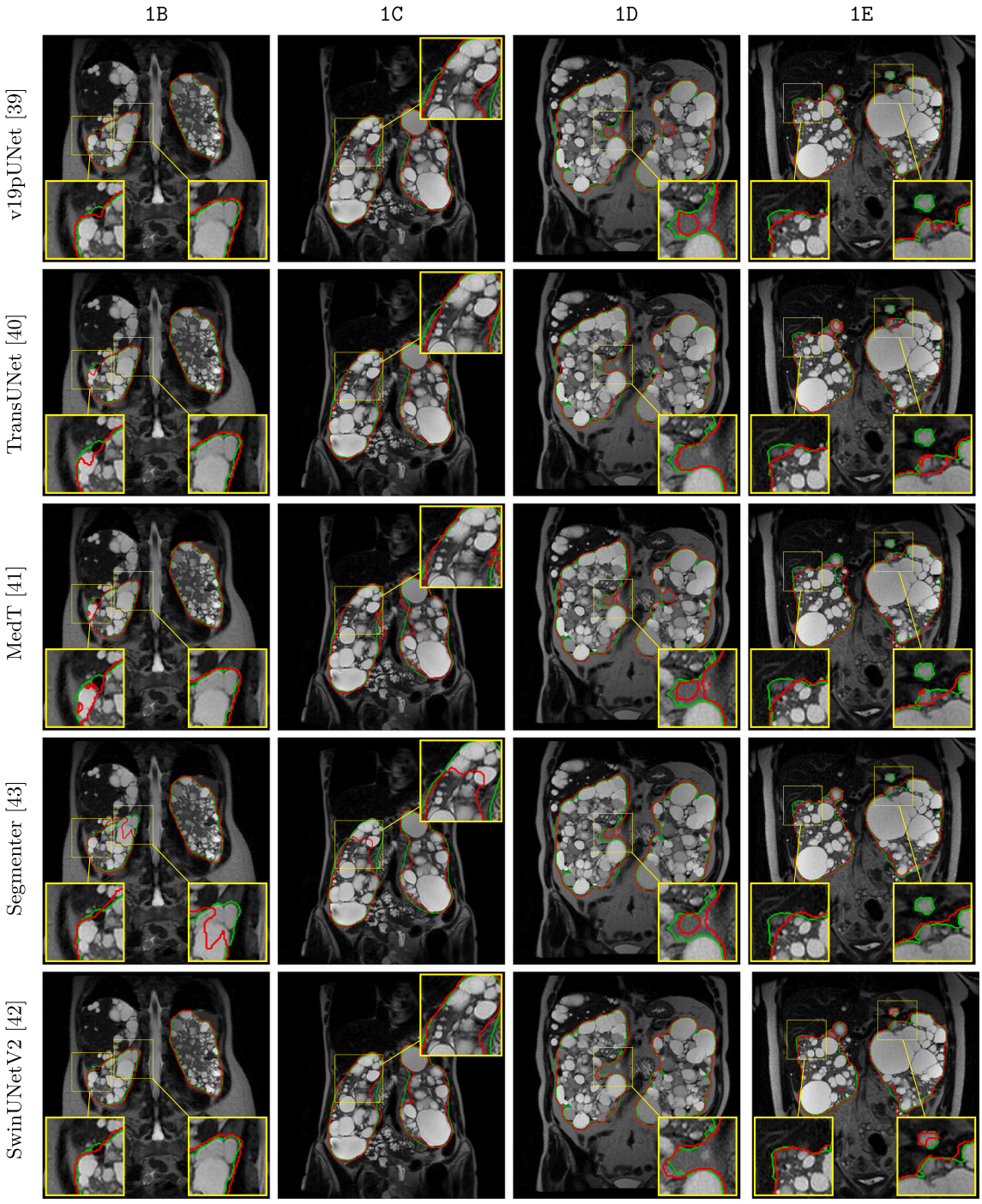

Fig. 13. Visual results using v19pUNet (Conze et al., 2020), TransUNet (Chen et al., 2021), MedT (Valanarasu et al., 2021), Segmenter (Strudel et al., 2021) and SwinUNetV2 (Liu et al., 2022) trained with our dual-task learning scheme (DT). The Mayo grading scale is covered from 1B to 1E. Ground truth and predicted contours are in green and red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Ultimately, our contributions could contribute to better perform patient stratification, treatment planning and progression monitoring.

CRediT authorship contribution statement

Pierre-Henri Conze: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Gustavo Andrade-Miranda:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation. **Yannick Le Meur:** Supervision, Project administration, Funding acquisition, Formal analysis, Conceptualization. **Emilie Cornec-Le Gall:** Writing – original draft, Supervision, Project administration, Funding acquisition, Formal analysis, Data curation, Conceptualization. **François Rousseau:** Writing – original draft, Supervision, Methodology, Investigation, Conceptualization.

Declaration of competing interest

None of the other authors of this manuscript have any financial or personal relationships with other people or organizations that could inappropriately influence and bias this work.

Data availability

The authors do not have permission to share data.

Acknowledgments

The authors acknowledge Christelle Ratajczak, Christelle Guillerm-Regost, Océane Pierry and Margaux Delaporte from University Hospital of Brest as well as Félix Renard and all Genkyst study investigators for data sharing and fruitful discussions.

References

Andrade-Miranda, G., Jaouen, V., Tankyevych, O., Le Rest, C.C., Visvikis, D., Conze, P.-H., 2023. Multi-modal medical transformers: A meta-analysis for medical image segmentation in oncology. Comput. Med. Imaging Graph. 110, 102308.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495.

Bevilacqua, V., Brunetti, A., Cascarano, G.D., Guerriero, A., Pesce, F., Moschetta, M., Gesualdo, L., 2019. A comparison between two semantic deep learning frameworks for the autosomal dominant polycystic kidney disease segmentation based on magnetic resonance images. BMC Med. Inform. Decis. Mak. 19 (9), 1–12.

Boutillon, A., Conze, P.-H., Pons, C., Burdin, V., Borotikar, B., 2022. Generalizable multi-task, multi-domain deep segmentation of sparse pediatric imaging datasets via multi-scale contrastive regularization and multi-joint anatomical priors. Med. Image Anal. 81, 102556.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229.

Chapman, A.B., Devuyst, O., Eckardt, K.-U., Gansevoort, R.T., Harris, T., Horie, S., Kasiske, B.L., Odland, D., Pei, Y., Perrone, R.D., et al., 2015. Autosomal-dominant polycystic kidney disease (ADPKD): executive summary from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. Kidney Int. 88 (1), 17–27.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Cirrincione, G., Cannata, S., Cicceri, G., Prinzi, F., Currieri, T., Lovino, M., Militello, C., Pasero, E., Vitabile, S., 2023. Transformer-based approach to melanoma detection. Sensors 23 (12), 5677.

Conze, P.-H., Andrade-Miranda, G., Singh, V.K., Jaouen, V., Visvikis, D., 2023. Current and emerging trends in medical image segmentation with deep learning. IEEE Trans. Radiat. Plasma Med. Sci..

Conze, P.-H., Brochard, S., Burdin, V., Sheehan, F.T., Pons, C., 2020. Healthy versus pathological learning transferability in shoulder muscle MRI segmentation using deep convolutional encoder-decoders. Comput. Med. Imaging Graph. 83, 101733.

Conze, P.-H., Kavur, A.E., Cornec-Le Gall, E., Gezer, N.S., Le Meur, Y., Selver, M.A., Rousseau, F., 2021. Abdominal multi-organ segmentation with cascaded convolutional and adversarial deep networks. Artif. Intell. Med. 117, 102109.

Cornec-Le Gall, E., Alam, A., Perrone, R.D., 2019. Autosomal dominant polycystic kidney disease. Lancet 393 (10174), 919–935.

Cornec-Le Gall, E., Torres, V.E., Harris, P.C., 2018. Genetic complexity of autosomal dominant polycystic kidney and liver diseases. J. Am. Soc. Nephrol. 29 (1), 13–23.

Daum, V., Helbig, H., Janka, R., Eckardt, K., Zeltner, R., 2007. In: Hornegger, J., et al. (Eds.), Quantitative Measurement of Kidney and Cyst Sizes in Patients with Autosomal Dominant Polycystic Kidney Disease (ADPKD). pp. 111–115.

Dhamija, T., Gupta, A., Gupta, S., et al., 2023. Semantic segmentation in medical images through transfused convolution and transformer networks. Appl. Intell. 53, 1132–1148.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Duran, A., Dussert, G., Rouvière, O., Jaouen, T., Jodoin, P.-M., Lartizien, C., 2022. ProstAttention-Net: A deep attention model for prostate cancer segmentation by aggressiveness in MRI scans. Med. Image Anal. 77.

Goel, A., Shih, G., Riyahi, S., Jeph, S., Dev, H., Hu, R., Romano, D., Teichman, K., Blumenfeld, J.D., Barash, I., et al., 2022. Deployed deep learning kidney segmentation for polycystic kidney disease MRI. Radiol. Artif. Intell.

Grantham, J.J., Torres, V.E., Chapman, A.B., Guay-Woodford, L.M., Bae, K.T., King, Jr., B.F., Wetzel, L.H., Baumgarten, D.A., Kenney, P.J., Harris, P.C., et al., 2006. Volume progression in polycystic kidney disease. N. Engl. J. Med. 354 (20), 2122–2130.

Guo, J., Odu, A., Pedrosa, I., 2022. Deep learning kidney segmentation with very limited training data using a cascaded convolution neural network. PLoS One 17 (5).

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., Xu, D., 2022a. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D., 2022b. UNETR: Transformers for 3D medical image segmentation. In: IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 272–284.

Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., et al., 2021. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge. Med. Image Anal. 67, 101821.

Higashihara, E., Nutahara, K., Okegawa, T., Shishido, T., Tanbo, M., Kobayasi, K., Nitadori, T., 2014. Kidney volume and function in autosomal dominant polycystic kidney disease. Clin. Exp. Nephrol. 18, 157–165.

Higashihara, E., Nutahara, K., Okegawa, T., Tanbo, M., Hara, H., Miyazaki, I., Kobayasi, K., Nitatori, T., 2015. Kidney volume estimations with ellipsoid equations by magnetic resonance imaging in autosomal dominant polycystic kidney disease. Nephron 129 (4), 253–262.

Jun, E., Jeong, S., Heo, D.-W., Suk, H.-I., 2021. Medical transformer: Universal brain encoder for 3D MRI analysis. arXiv preprint arXiv:2104.13633.

Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.-H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., et al., 2021. CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation. Med. Image Anal. 69, 101950.

Kim, Y., Ge, Y., Tao, C., Zhu, J., Chapman, A.B., Torres, V.E., Alan, S., Mrug, M., Bennett, W.M., Flessner, M.F., et al., 2016. Automated segmentation of kidneys from MR images in patients with autosomal dominant polycystic kidney disease. Clin. J. Am. Soc. Nephrol. 11 (4), 576–584.

Kline, T.L., Korfiatis, P., Edwards, M.E., Blais, J.D., Czerwiec, F.S., Harris, P.C., King, B.F., Torres, V.E., Erickson, B.J., 2017. Performance of an artificial multi-observer deep neural network for fully automated segmentation of polycystic kidneys. J. Dig. Imaging 30 (4), 442–448.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al., 2022. Swin transformer v2: Scaling up capacity and resolution. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12009–12019.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Mignani, R., Corsi, C., De Marco, M., Caiani, E.G., Santucci, G., Cavagna, E., Severi, S., Cagnoli, L., 2011. Assessment of kidney volume in polycystic kidney disease using magnetic resonance imaging without contrast medium. Am. J. Nephrol. 33 (2), 176–184.

Racimora, D., Vivier, P.-H., Chandarana, H., Rusinek, H., 2010. Segmentation of polycystic kidneys from MR images. In: Medical Imaging 2010: Computer-Aided Diagnosis, vol. 7624, pp. 548–558.

Raj, A., Tollens, F., Hansen, L., Golla, A.-K., Schad, L.R., Nörenberg, D., Zöllner, F.G., 2022. Deep learning-based total kidney volume segmentation in autosomal dominant polycystic kidney disease using attention, cosine loss, and sharpness aware minimization. Diagnostics 12 (5), 1159.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention. pp. 234–241.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115.

Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H., 2022. Transformers in medical imaging: A survey. arXiv preprint arXiv:2201.09873.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Strudel, R., Garcia, R., Laptev, I., Schmid, C., 2021. Segmenter: Transformer for semantic segmentation. In: IEEE/CVF International Conference on Computer Vision. pp. 7262–7272.

Tan, M., Le, Q., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., 2021. Training data-efficient image transformers and distillation through attention. In: International Conference on Machine Learning.

Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M., 2021. Medical transformer: Gated axial-attention for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 36–46.

van Gastel, M.D., Edwards, M.E., Torres, V.E., Erickson, B.J., Gansevoort, R.T., Kline, T.L., 2019. Automatic measurement of kidney and liver volumes from MR images of patients affected by autosomal dominant polycystic kidney disease. J. Am. Soc. Nephrol..

Yan, Y., Conze, P.-H., Lamard, M., Quellec, G., Cochener, B., Coatrieux, G., 2021. Towards improved breast mass detection using dual-view mammogram matching. Med. Image Anal. 71, 102083.

Zhao, Y., Wang, X., Che, T., Bao, G., Li, S., 2022. Multi-task deep learning for medical image computing and analysis: A review. Comput. Biol. Med. 106496.

Zöllner, F.G., Kociński, M., Hansen, L., Golla, A.-K., Trbalić, A.Š., Lundervold, A., Materka, A., Rogelj, P., 2021. Kidney segmentation in renal magnetic resonance imaging-current status and prospects. IEEE Access 9, 71577–71605.