Computer aided diagnosis of congenital abnormalities of the kidney and urinary tract in children based on ultrasound imaging data by integrating texture image features and deep transfer learning image features

Feature extraction methods

1. Image normalization

We normalized ultrasound kidney images of different subjects using as following. First, the orientation of kidneys in ultrasound images was estimated based on ellipse fitting, including the ellipse's major axis, minor axis, and the orientation θ between the major axis and X-axis, as illustrated in Figure S1. Second, based on the estimated ellipse information, each ultrasound kidney image was reoriented along the major axis based on a rotation matrix estimated from the ellipse fitting, as formulated by Eq. (1)

$$T = \begin{bmatrix} cos(\theta) & sin(\theta) & 0 \\ -sin(\theta) & cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (1)$$

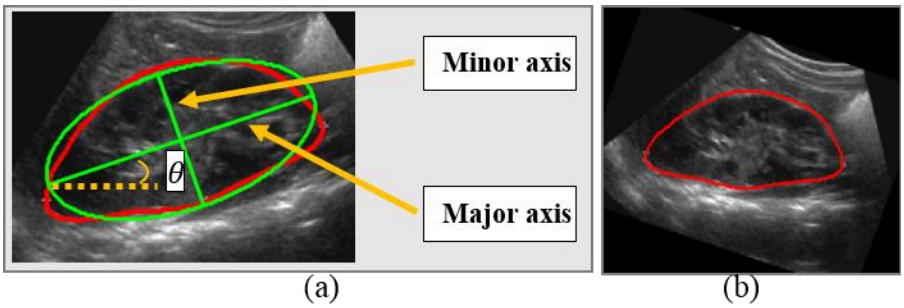

Figure S1. Kidney image normalization based on ellipse estimation of kidney boundary. (a) Estimation of an ellipse of the kidney boundary; (b) Reorientation of the kidney image based on the estimated ellipse of kidney.

2. Transfer learning based imaging feature extraction

A pre-trained CNN model (imagenet-caffe-alex)[1] was adopted from MatConvNet[2] to extract deep learning features from the kidney images in a transfer learning setting. The imagenet-caffe-alex model was trained on 1.2 million 3-channel images of the ImageNet

LSVRC-2010 for classifying images into 1000 different classes. The model has 5 convolutional layers, followed by max-pooling layers and 3 fully-connected layers with a final 1000-way softmax output. Figure S2 shows architecture of the imagenet-caffe-alex model, including 5 layers of convolutional neural networks, 3 fully connected layers, and a softmax output layer.

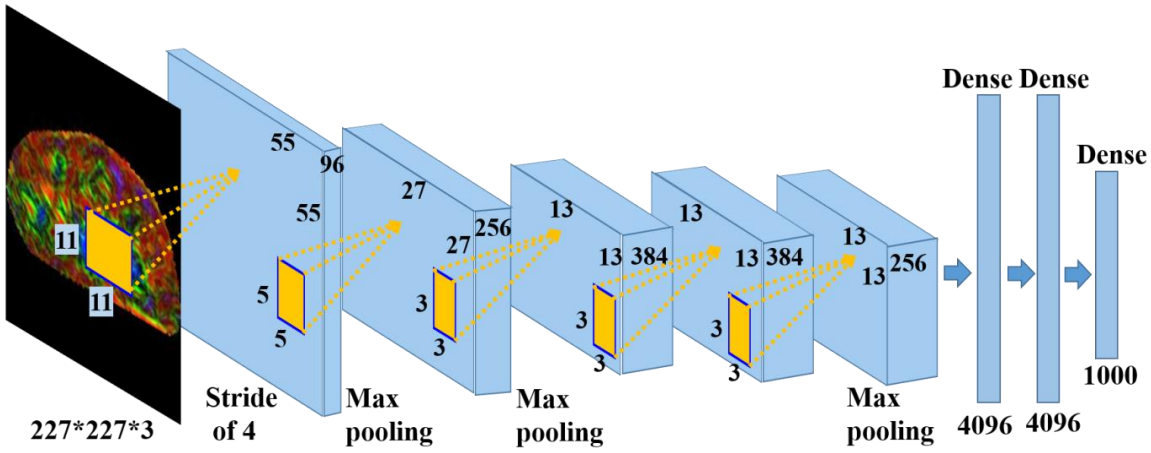

Figure S2. Architecture of the imagenet-caffe-alex model, including 5 layers of convolutional (conv) neural networks, 3 fully connected (fc) layers, and a softmax output layer. Rectified linear unit (relu), normalization (norm), pooling (pool), and dropout operators are used in different layers.

Since the imagenet-caffe-alex network needs 3-channel (RGB) images as its input, we generated 3-channel images from the kidney ultrasound images by adopting the original kidney image $f_I(x,y)$, a gradient feature map $f_G(x,y)$, and a distanced transform map $f_D(x,y)$ with x and y be coordinates of ultrasound image pixels as 3 channels, as illustrated in Figure S3.

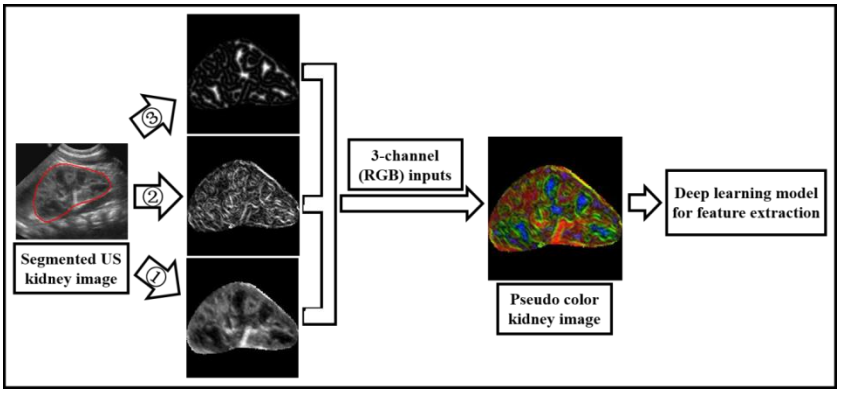

Figure S3. Feature extraction by transfer learning from pseudo color images with ① image intensity map, ② gradient feature map, and ③ distanced transform feature map as RGB-channels respectively.

Particularly, the image intensity map $f_I(x,y)$ was obtained from the original ultrasound kidney image intensity values after normalized into [0, 255]. Based on the image intensity map $f_I(x,y)$, a gradient feature map $f_G(x,y)$ was computed as:

$$f_G(x,y) = \frac{g(x,y)}{f_I(x,y)} = \frac{\sqrt{g_x^2(x,y)+g_y^2(x,y)}}{f_I(x,y)}, \quad (2)$$

where $g_x(x,y) = \big(f_I(x+1,y) - f_I(x-1,y)\big)/2$, $g_y(x,y) = \big(f_I(x,y+1) - f_I(x,y-1)\big)/2$, and x and y are coordinates of pixels. A distance transform feature map for characterizing the distance of each pixel to its nearest element in an edge map[3] was computed using the VLfeat toolbox [4] as

$$f_D(x,y) = \min_{x',y' \in \{f_{edge}(x',y')>0\}} (x-x')^2 + (y-y')^2, \quad (3)$$

where $f_{edge}(x',y')$ is an edge image of the original image obtained by Canny edge detector[3], and $(x-x')^2 + (y-y')^2$ is the distance between pixel (x,y) and (x',y').

The image intensity map $f_I(x,y)$, gradient feature map $f_G(x,y)$, and distanced transform feature map $f_D(x,y)$ of each kidney scan were finally normalized as

$$f(x,y) = \frac{f(x,y) - mean(f(x,y))}{std(f(x,y))}, f(x,y) \in \{f_I(x,y), f_G(x,y), f_D(x,y)\}. \quad (4)$$

The normalization was only applied to pixels within the kidney region and the feature values were finally linearly scaled into [0, 255].

After cropped based on a bounding box of the reoriented kidney mask, the 3 feature maps were resized to have the same size as the images used in the imagenet-caffe-alex net. Particularly, given the input image size of the deep learning model, $[N_0, N_0]$, and a ultrasound kidney image's $ratio = minor\ axis/major\ axis$ of its estimated ellipse, the cropped feature images were resized to $[N_r, N_0]$, where N_0 is the length of the major axes, and $N_r = ratio \times N_0$ is to guarantee the ratio of kidney fitting shape. The resized feature maps of $[N_r, N_0]$ were finally padded with zeroes to have an image size of $[N_0, N_0]$. For the imagenet-caffe-alex model, N_0=227.

Finally, the image intensity map, gradient feature map, and distanced transform feature map are used as R-channel, G-channel, and B-channel respectively to form a 3-channel image so that the imagenet-caffe-alex model could be adopted to extract deep learning features. Figure S4 shows 4 randomly selected ultrasound images and their corresponding RGB channels, as well as their pseudo color images. Since a permute of

the order of feature maps generates different 3-channel images, we evaluated how the classification performance changed with the order of feature maps.

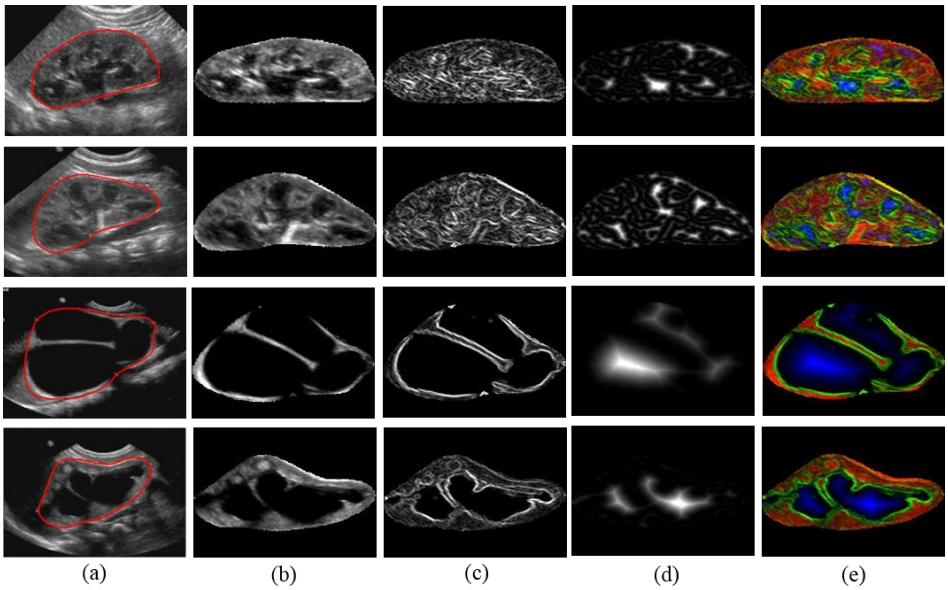

Figure S4. Four example ultrasound kidney images, feature maps, and pseudo color images. (a) ultrasound images with kidney contours in red; (b) image intensity maps $f_I(x,y)$; (c) gradient feature maps $f_G(x,y)$; (d) distanced transform feature maps $f_D(x,y)$, and (e) pseudo color images.

3. Conventional imaging feature extraction

Conventional image features were also extracted from the image intensity map $f_I(x,y)$, including geometrical features[5] and histogram of oriented gradients (HOG) features[6] of the kidneys.

Geometrical feature extraction

The geometrical features included shape-related measures and hole-related measures of the kidneys. The shape-related features were computed based on major axis and minor axis estimated based on the ellipse fitting results, defined as $\mathrm{V_{shape}} = [\mathrm{L_1}, \mathrm{L_2}, \mathrm{L_1/L_2}, \mathrm{L_1} \times \mathrm{L_2}, \mathrm{L_1} + \mathrm{L_2}, \mathrm{L_1^2} + \mathrm{L_2^2}, \mathrm{L_1} - \mathrm{L_2}, \mathrm{L_1^2} - \mathrm{L_2^2}]$, where L_1 and L_2 are lengths of the major and minor axes of the kidney. The hole-related features were ratios of areas of black holes inside kidney to the whole kidney region, surrogate measures of renal parenchymal area[5]. Since no suitable threshold was available for segmenting holes in all kidney images, we uniformly set 10 thresholds of $[3{:}\,3{:}\,30]$ to segment black holes, and obtained hole-related

features $\boldsymbol{V}_{hole} = [ratio_1, \cdots, ratio_{10}]$. Finally, from each kidney image, we obtained a set of geometric features $\boldsymbol{V}_{\boldsymbol{geometric}} = [\boldsymbol{V}_{shape}, \boldsymbol{V}_{hole}]$ with 18 elements.

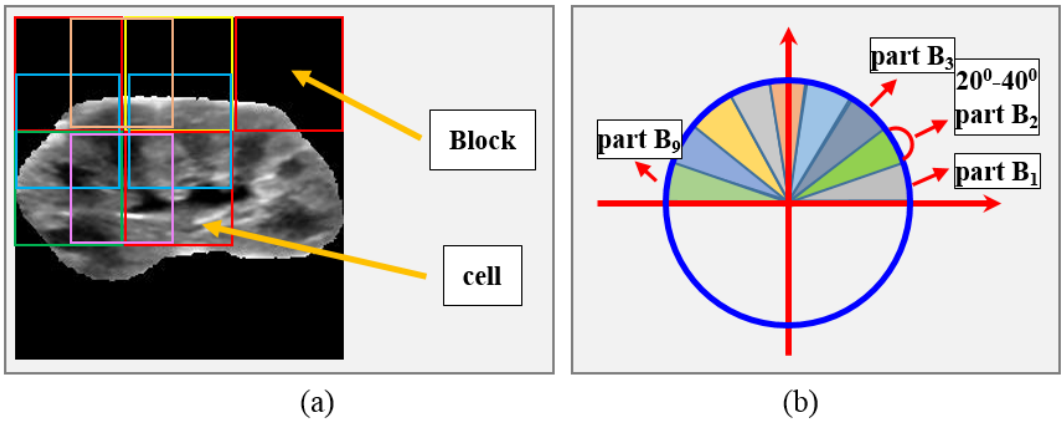

Figure S5. HOG feature extraction. (a) Cells and blocks. (b) orientation-based histogram.

HOG feature extraction

The HOG feature extraction method typically decomposes an image into small squared cells, computes histogram of oriented gradients in each cell, normalizes the result using a block-wise pattern, and yields a descriptor for each cell, as illustrated in Figure S5. Particularly, the magnitude and the orientation of the gradient of image pixel at (x, y) were computed as $|G| = \sqrt{I_x^2 + I_y^2}$ and $\theta = arctan(I_x/I_y)$, where $I_x = f_I(x, y) * D_x$ and $I_y = f_I(x, y) * D_y$ are x and y derivatives, computed by convolution operations based on 1-D horizontal and vertical derivative masks $D_x = [-1 \quad 0 \quad 1]$ and $D_y = [-1 \quad 0 \quad 1]^T$. By splitting the image intensity map into $N_{cell} \times N_{cell}$ cells, as illustrated in Figure S5(a), an orientation-based histogram was estimated for each cell based on the gradient orientation information of pixels within the cell as illustrated in Figure S5(b), and each bin was weighted by their voxels' gradient magnitude information.

In our study, 9 orientation bins over 0-180 degrees were adopted to estimate the histogram of oriented gradient for each cell as recommended in a previous study[7]. Based on the cell definition, blocks were built by grouping 4 neighbored cells into a larger overlapped block with size $2N_{cell} \times 2N_{cell}$, and block features v were computed by integrating the 4 neighbored cells' features. The block features were normalized as $\frac{v}{\sqrt{\|v\|_2^2 + e^2}}$, where v is a block feature vector before normalization, $\|\cdot\|_2$ is l^2norm, and e is a

small constant. By sliding the block on the image with a stride $N_{stride} = N_{cell}$, a set of features were obtained by concatenating all overlapped block features together. We utilized the VLFeat toolbox [4] to compute the HOG features with default parameters except that the cell size was set to $N_{cell} = N_0 \,/\, 10$, where N_0 is the size of the input images to the deep learning model. The block feature v for each block contained 31 elements, and the total number of features was 3100.

4. Support vector machine classification

A L2-regularized L1-loss support vector machine (SVM) classifier[8] was utilized to build classifiers based on the extracted image features by optimizing

$$\min_{w} \frac{1}{2}\vec{w}^T\vec{w} + C\sum_{i=1}^{l}\left(\max\left(0,1 - y_i\vec{w}^T\vec{f}_i\right)\right) \quad (5)$$

where $\vec{f}_i$ is the feature vector of the i^{th} ultrasound kidney image, $\vec{w}$ is the weighting vector to be learned from training data.

The L2-regularized L1-loss SVM optimization problem can be solved by using a dual coordinate descent method. Particularly, a publicly available software package LIBLINEAR with its default parameters can be utilized to build the SVM classifiers[8]. Once we obtained the weighting vector $\vec{w}$, the category of the i^{th} ultrasound kidney image can be estimated as

$$L_i = sgn\left(\vec{w}^T\vec{f}_i\right) . \quad (6)$$

5. Classification performance of CNN features extracted from different pseudo color images

To obtain pseudo color images with the most discriminative information for the transfer learning, 7 different pseudo color images were generated with permuted orders of the feature maps including the image intensity map $f_I(x,y)$, the gradient feature map $f_G(x,y)$, and the distanced transform feature map $f_D(x,y)$ as well as 3 duplication of the image intensity map $f_I(x,y)$ as the RGB-channels and, as summarized in Table S1.

Table S1. Pseudo color images generated for transfer learning

	C1	C2	C3	C4	C5	C6	C7
R-channel	$f_I(x,y)$	$f_I(x,y)$	$f_D(x,y)$	$f_D(x,y)$	$f_G(x,y)$	$f_G(x,y)$	$f_I(x,y)$
G-channel	$f_G(x,y)$	$f_D(x,y)$	$f_G(x,y)$	$f_I(x,y)$	$f_I(x,y)$	$f_D(x,y)$	$f_I(x,y)$
B-channel	$f_D(x,y)$	$f_G(x,y)$	$f_I(x,y)$	$f_G(x,y)$	$f_D(x,y)$	$f_I(x,y)$	$f_I(x,y)$

Table S2 summarizes their classification performance estimated based on 100 runs of 10 fold cross-validation of SVM classifiers built upon the transfer learning image features learned from different 3-channel images, including accuracy, specificity, sensitivity, and AUC. Overall, the pseudo color images generated by scheme C5 (R-channel: gradient map, G-channel: US image, B-channel: distance map) yielded the best performance. The scheme C5 was finally adopted in this study. Most of the classification models built on the transfer learning based features had better classification performance than the best classification model built on the conventional imaging features (HOG + Geometrical image features), except the classification model built on the transfer learning features extracted from pseudo color images of C4 and C7.

Figure S6 shows the ROC curves of one run of 10-fold cross-validation for different pseudo color images.

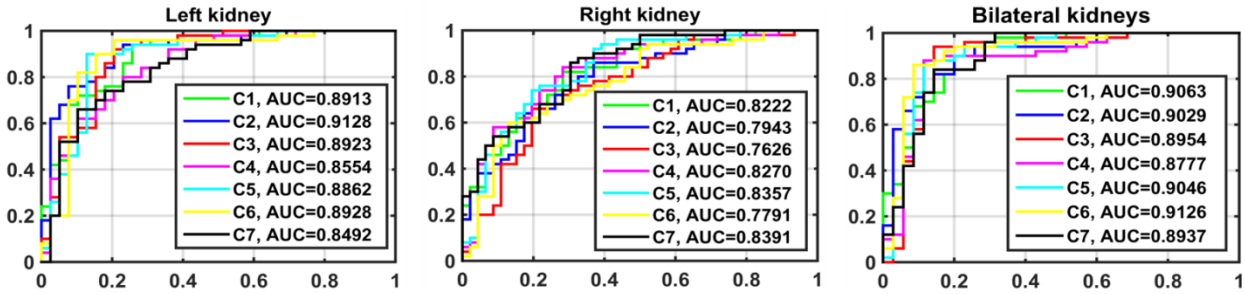

Figure S6. ROC curves of 7 different pseudo color images, estimated based on one run of 10-fold cross-validation.

Table S2. Classification performance of different pseudo color images, estimated based on 100 runs of 10-fold cross-validation (mean±std).

	C1	C2	C3	C4	C5	C6	C7
				Accuracy			
Left	*0.80±1.3e-2	*0.81±1.6e-2	*0.84±1.5e-2	0.77±2.1e-2	***0.85±1.8e-2**	*0.86±1.6e-2	0.78±2.6e-2
Right	0.74±2.0e-2	0.71±2.0e-2	0.69±2.4e-2	*0.77±1.4e-2	**0.75±2.0e-2**	0.69±1.7e-2	0.76±2.4e-2
Bilateral	*0.82±2.0e-2	*0.83±1.7e-2	*0.86±1.4e-2	*0.83±2.0e-2	***0.85±1.2e-2**	*0.85±1.6e-2	0.73±1.7e-2
				AUC			
Left	*0.89±1.2e-2	*0.90±1.0e-2	*0.89±0.9e-2	0.85±1.2e-2	***0.88±0.8e-2**	*0.89±0.8e-2	0.86±1.2e-2
Right	0.81±1.2e-2	0.79±1.3e-2	0.75±1.4e-2	0.81±1.0e-2	**0.83±1.3e-2**	0.78±1.3e-2	0.85±1.2e-2
Bilateral	*0.90±0.9e-2	*0.89±0.9e-2	*0.88±0.7e-2	0.86±0.8e-2	***0.89±0.8e-2**	*0.90±0.7e-2	0.89±1.0e-2
				Specificity			
Left	*0.78±1.6e-2	*0.79±2.1e-2	*0.79±1.4e-2	*0.75±2.7e-2	***0.80±2.6e-2**	*0.80±2.3e-2	0.82±3.1e-2
Right	0.63±2.5e-2	*0.65±3.0e-2	*0.68±2.8e-2	*0.77±1.8e-2	***0.66±2.6e-2**	*0.67±3.0e-2	0.71±3.0e-2
Bilateral	*0.84±2.1e-2	0.81±2.5e-2	*0.87±1.6e-2	*0.82±2.9e-2	***0.82±1.4e-2**	*0.89±2.2e-2	0.88±0.7e-2
				Sensitivity			
Left	0.82±2.0e-2	0.84±2.4e-2	*0.88±2.3e-2	0.78±3.2e-2	***0.89±2.2e-2**	*0.91±2.2e-2	0.74±3.7e-2
Right	0.84±2.6e-2	0.78±2.8e-2	0.70±3.2e-2	0.76±2.3e-2	**0.83±2.8e-2**	*0.70±2.2e-2	0.80±3.4e-2
Bilateral	*0.80±2.9e-2	*0.84±2.2e-2	*0.86±2.2e-2	*0.84±2.5e-2	***0.86±1.9e-2**	*0.82±2.1e-2	0.62±2.8e-2

* CNN feature better than conventional features (HOG + Geometrical); Wilcoxon signed-rank test: p<=0.001

References

1. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems 25 (NIPS 2012).* 2012:1097-1105.

2. MatConvNet team. MatConvNet: CNNs for MATLAB. Vol 20172017.

3. Felzenszwalb PF, Huttenlocher DP. Distance transforms of sampled functions. *Techinical Report, Cornell University.* 2004.

4. Vedaldi A, Fulkerson B. VLFeat: An open and portable library of computer vision algorithms. *in Proceedings of the 18th Annual ACM International Conference on Multimedia*. Firenze, Italy2010:1469-1472.

5. Pulido JE, Furth SL, Zderic SA, Canning DA, Tasian GE. Renal parenchymal area and risk of ESRD in boys with posterior urethral valves. *Clinical journal of the American Society of Nephrology : CJASN.* 2014;9:499-505.

6. Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2005;II:886-893.

7. Dalal N, Triggs B. Histograms of oriented gradients for human detection. *In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA.* 2005;1:886-893.

8. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research.* 2008;9:1871-1874.