TABLE I: Comparison robustness performances among different data corruption in the lung segmentation task on JSRT dataset using UNet [27] as baseline. Dice is utilized as the evaluation metric.

Corruption	UNet (—)	UNet (+ VAE)	UNet (+**Ours**)	**SWAE**-UNet (—)	**SWAE**-UNet (+**Ours**)
Original	94.18 ± 0.2	95.43 ± 0.2	95.41 ± 0.1	95.07 ± 0.1	**96.56 ± 0.2**
Gauss. Noise	74.02 ± 0.4	75.70 ± 0.2	94.96 ± 0.1	86.26 ± 0.2	**95.30 ± 0.1**
Shot Noise	65.85 ± 0.1	74.14 ± 0.3	**94.53 ± 0.3**	78.52 ± 0.1	94.50 ± 0.3
Impulse Noise	72.78 ± 0.2	66.80 ± 0.1	**95.01 ± 0.1**	87.21 ± 0.5	94.81 ± 0.3
Speckle Noise	72.12 ± 0.4	74.34 ± 0.3	93.78 ± 0.3	82.44 ± 0.3	**94.67 ± 0.2**
Poisson Noise	76.33 ± 0.5	75.89 ± 0.1	**95.02 ± 0.2**	87.11 ± 0.1	94.83 ± 0.5
Dropout	77.50 ± 0.3	78.56 ± 0.4	94.60 ± 0.1	74.63 ± 0.2	94.83 ± 0.2
Gauss. Blur	93.38 ± 0.1	87.54 ± 0.1	95.26 ± 0.3	92.76 ± 0.4	**96.03 ± 0.2**
Glass Blur	94.13 ± 0.2	86.58 ± 0.3	94.86 ± 0.5	93.76 ± 0.2	**96.07 ± 0.4**
Defocus Blur	92.89 ± 0.3	95.44 ± 0.2	93.22 ± 0.1	86.96 ± 0.1	**95.87 ± 0.3**
Motion Blur	87.46 ± 0.2	83.77 ± 0.3	89.40 ± 0.4	87.12 ± 0.5	**90.06 ± 0.1**
Zoom Blur	**85.25 ± 0.5**	79.42 ± 0.2	82.79 ± 0.1	83.56 ± 0.3	84.66 ± 0.4
Fog	58.17 ± 0.6	64.72 ± 0.1	**70.45 ± 0.4**	62.93 ± 0.2	70.40 ± 0.2
Contrast	03.82 ± 0.2	37.59 ± 0.2	54.29 ± 0.2	03.06 ± 0.5	**65.80 ± 0.1**
Brightness	05.78 ± 0.3	42.29 ± 0.1	71.03 ± 0.3	09.54 ± 0.2	**74.69 ± 0.2**
Saturate	94.72 ± 0.5	88.15 ± 0.2	94. 97 ± 0.3	93.94 ± 0.1	**96.18 ± 0.3**
JpegComp.	94.01 ± 0.1	88.14 ± 0.3	95. 32 ± 0.1	93.75 ± 0.4	**96.23 ± 0.4**
Elastic Trans.	93.08 ± 0.3	93.05 ± 0.2	95. 49 ± 0.3	93.45 ± 0.1	**95.51 ± 0.2**
Avgerage	67.61	75.51	88.66	76.66	**90.02**
↓	26.57	19.92	06.75	18.41	**06.54**

↓ denotes that compared with the clean data (original) the average dice descent by.

TABLE II: Comparison robustness performances among different data corruption in the lung segmentation task on JSRT dataset using PSPNet [30] as baseline. Dice is utilized as the evaluation metric.

Corruption	PSPNet (—)	PSPNet (+VAE)	PSPNet (+ImageNet)	PSPNet (+Jigsaw)	PSPNet (+MoCo)	PSPNet (+**Ours**)	**SWAE**-PSPNe (—)	**SWAE**-PSPNet (+MoCo)	**SWAE**-PSPNet (+**Ours**)
Original	95.34 ± 0.2	94.26 ± 0.3	96.54 ± 0.1	96.32 ± 0.3	96.50 ± 0.1	96.24 ± 0.2	95.52 ± 0.3	96.80 ± 0.1	**97.19 ± 0.1**
Gauss. Noise	84.28 ± 0.4	89.71 ± 0.2	80.97 ± 0.1	78.97 ± 0.1	87.62 ± 0.3	95.45 ± 0.2	85.35 ± 0.4	88.77 ± 0.2	**95.61 ± 0.1**
Shot Noise	88.35 ± 0.2	88.03 ± 0.1	79.51 ± 0.2	78.67 ± 0.1	87.56 ± 0.4	95.00 ± 0.2	88.53 ± 0.3	89.53 ± 0.1	**95.80 ± 0.2**
Impulse Noise	87.56 ± 0.4	89.46 ± 0.1	78.92 ± 0.4	81.40 ± 0.3	84.10 ± 0.2	95.80 ± 0.4	86.82 ± 0.3	87.08 ± 0.4	**95.89 ± 0.2**
Speckle Noise	88.64 ± 0.1	88.93 ± 0.3	79.13 ± 0.2	83.62 ± 0.1	85.63 ± 0.2	95.86 ± 0.2	88.61 ± 0.2	90.00 ± 0.2	**95.92 ± 0.1**
Poisson Noise	82.38 ± 0.3	89.11 ± 0.2	82.89 ± 0.2	78.17 ± 0.2	90.49 ± 0.3	94.51 ± 0.4	85.23 ± 0.4	91.03 ± 0.3	**95.01 ± 0.3**
Dropout	87.59 ± 0.1	86.96 ± 0.3	70.69 ± 0.2	73.19 ± 0.5	86.79 ± 0.1	94.94 ± 0.2	86.51 ± 0.1	86.06 ± 0.4	**95.47 ± 0.4**
Gauss. Blur	94.72 ± 0.2	93.23 ± 0.2	93.70 ± 0.3	93.91 ± 0.2	94.69 ± 0.3	**96.42 ± 0.1**	95.58 ± 0.1	95.54 ± 0.1	**96.42 ± 0.2**
Glass Blur	94.77 ± 0.3	93.88 ± 0.2	91.24± 0.1	92.83 ± 0.3	93.32 ± 0.2	**96.55 ± 0.2**	95.48 ± 0.4	95.69 ± 0.4	96.51 ± 0.1
Defocus Blur	94.87 ± 0.1	93.40 ± 0.1	93.94 ± 0.3	94.03 ± 0.3	94.95 ± 0.2	96.48 ± 0.3	95.78 ± 0.1	95.58 ± 0.3	**96.53 ± 0.2**
Motion Blur	88.97 ± 0.1	86.50 ± 0.3	88.01 ± 0.2	90.52 ± 0.4	91.76 ± 0.3	90.01 ± 0.2	89.62 ± 0.1	**92.74 ± 0.2**	90.26 ± 0.1
Zoom Blur	82.65 ± 0.4	83.87 ± 0.2	85.88 ± 0.3	84.91 ± 0.1	89.11 ± 0.2	84.47 ± 0.1	85.44 ± 0.2	**89.34 ± 0.1**	85.23 ± 0.2
Fog	58.48 ± 0.2	61.80 ± 0.1	77.58 ± 0.4	76.30 ± 0.2	**79.77 ± 0.2**	74.50 ± 0.3	61.68 ± 0.3	79.55 ± 0.2	76.48 ± 0.1
Contrast	43.77 ± 0.5	60.73 ± 0.3	59.07 ± 0.2	56.59 ± 0.4	65.55 ± 0.4	65.37 ± 0.3	55.71 ± 0.4	65.47 ± 0.2	**67.74 ± 0.3**
Brightness	00.75 ± 0.1	72.73 ± 0.2	81.99 ± 0.2	82.08 ± 0.3	82.93 ± 0.3	79.89 ± 0.1	01.27 ± 0.1	**83.78 ± 0.1**	81.95 ± 0.1
Saturate	95.34 ± 0.2	94.26 ± 0.1	96.52 ± 0.1	96.32 ± 0.1	96.47 ± 0.1	96.18 ± 0.1	96.32 ± 0.2	96.42 ± 0.3	**97.11 ± 0.2**
JpegComp.	95.16 ± 0.1	94.09 ± 0.1	91.22 ± 0.4	91.47 ± 0.3	92.88 ± 0.3	96.30 ± 0.3	96.11 ± 0.1	93.50 ± 0.3	**96.88 ± 0.3**
Elastic Trans.	95.15 ± 0.4	94.08 ± 0.1	94.07 ± 0.3	95.32 ± 0.1	96.65 ± 0.2	96.55 ± 0.1	95.90 ± 0.3	96.27 ± 0.2	**96.74 ± 0.1**
Avgerage	80.20	85.93	83.84	84.01	88.25	90.65	81.76	89.20	**91.51**
↓	15.14	08.33	12.80	12.31	07.99	05.59	13.76	07.60	**05.68**

↓ denotes that compared with the clean data (original) the average dice descent by.

TABLE III: Comparison robustness performances among different data corruption in the lung segmentation task on SH dataset using UNet [27] as baseline. Dice is utilized as the evaluation metric.

Corruption	UNet (—)	UNet (+ VAE)	UNet (+**Ours**)	**SWAE**-UNet (—)	**SWAE**-UNet (+**Ours**)
Original	88.03 ± 0.2	86.66 ± 0.2	91.31 ± 0.2	89.34 ± 0.4	**93.17 ± 0.3**
Gauss. Noise	85.97 ± 0.4	86.47 ± 0.2	**90.50 ± 0.2**	86.59 ± 0.2	90.37 ± 0.1
Shot Noise	83.78 ± 0.3	84.93 ± 0.4	**90.61 ± 0.3**	83.09 ± 0.2	88.45 ± 0.2
Impulse Noise	86.96 ± 0.2	85.57 ± 0.1	90.18 ± 0.4	87.09 ± 0.3	**90.22 ± 0.3**
Speckle Noise	82.03 ± 0.2	82.91 ± 0.4	87.64 ± 0.2	81.31 ± 0.1	**88.81 ± 0.3**
Poisson Noise	84.03 ± 0.3	84.23 ± 0.4	87.73 ± 0.3	83.60 ± 0.1	**88.36 ± 0.1**
Dropout	88.30 ± 0.3	86.17 ± 0.2	**89.03 ± 0.3**	84.24 ± 0.2	88.99 ± 0.2
Gauss. Blur	85.07 ± 0.1	84.86 ± 0.2	89.47 ± 0.3	85.79 ± 0.2	**89.93 ± 0.2**
Glass Blur	85.25 ± 0.2	86.43 ± 0.4	90.76 ± 0.1	85.79 ± 0.1	**91.56 ± 0.2**
Defocus Blur	85.05 ± 0.1	86.66 ± 0.4	**90.90 ± 0.2**	86.53 ± 0.2	90.05 ± 0.4
Motion Blur	80.73 ± 0.2	81.62 ± 0.2	84.91 ± 0.5	82.45 ± 0.2	**85.56 ± 0.3**
Zoom Blur	69.43 ± 0.3	71.28 ± 0.1	77.68 ± 0.2	70.19 ± 0.1	**79.06 ± 0.4**
Fog	55.31 ± 0.4	51.22 ± 0.4	56.70 ± 0.2	53.25 ± 0.2	**56.98 ± 0.4**
Contrast	37.75 ± 0.2	35.45 ± 0.2	**43.19 ± 0.2**	35.85 ± 0.3	41.78 ± 0.1
Brightness	57.72 ± 0.4	61.48 ± 0.1	62.70 ± 0.2	**65.68 ± 0.3**	65.41 ± 0.1
Saturate	87.64 ± 0.4	88.93 ± 0.4	**91.72 ± 0.3**	86.90 ± 0.2	90.96 ± 0.1
JpegComp.	87.55 ± 0.4	88.44 ± 0.2	91.34 ± 0.2	87.72 ± 0.1	**91.76 ± 0.2**
Elastic Trans.	86.07 ± 0.2	86.47 ± 0.4	91.48 ± 0.2	86.82 ± 0.2	**92.95 ± 0.3**
Avgerage	78.16	78.42	**84.50**	78.41	83.01
↓	09.87	08.24	**06.81**	10.93	10.17

↓ denotes that compared with the clean data (original) the average dice descent by.

TABLE IV: Comparison robustness performances among different data corruption in the lung segmentation task on SH dataset using PSPNet [30] as baseline. Dice is utilized as the evaluation metric.

Corruption	PSPNet (—)	PSPNet (+VAE)	PSPNet (+ImageNet)	PSPNet (+Jigsaw)	PSPNet (+MoCo)	PSPNet (+**Ours**)	**SWAE**-PSPNet (—)	**SWAE**-PSPNet (+MoCo)	**SWAE**-PSPNet (+**Ours**)
Original	83.26 ± 0.3	91.97 ± 0.2	92.67 ± 0.2	93.95 ± 0.1	94.74 ± 0.1	92.68 ± 0.3	85.17 ± 0.3	**94.80 ± 0.2**	94.77 ± 0.2
Gauss. Noise	82.45 ± 0.2	83.43 ± 0.1	85.06 ± 0.2	86.99 ± 0.4	87.46 ± 0.2	85.93 ± 0.4	84.32 ± 0.1	90.38 ± 0.2	**92.03 ± 0.2**
Shot Noise	82.72 ± 0.1	76.10 ± 0.3	85.57 ± 0.3	85.38 ± 0.2	87.10 ± 0.4	84.96 ± 0.1	83.02 ± 0.3	89.95 ± 0.2	**91.00 ± 0.2**
Impulse Noise	81.46 ± 0.2	85.22 ± 0.3	85.61 ± 0.1	88.67 ± 0.4	85.48 ± 0.3	86.28 ± 0.2	83.09 ± 0.3	88.36 ± 0.1	**92.16 ± 0.3**
Speckle Noise	82.77 ± 0.2	83.72 ± 0.3	83.02 ± 0.4	83.08 ± 0.1	86.17 ± 0.3	84.20 ± 0.2	81.96 ± 0.3	86.33 ± 0.1	**90.00 ± 0.1**
Poisson Noise	79.89 ± 0.2	53.77 ± 0.4	88.43 ± 0.3	78.31 ± 0.2	89.63 ± 0.1	84.44 ± 0.3	83.88 ± 0.1	**91.65 ± 0.3**	90.88 ± 0.2
Dropout	79.48 ± 0.2	67.69 ± 0.2	77.57 ± 0.1	81.24 ± 0.3	76.35 ± 0.1	82.73 ± 0.2	75.57 ± 0.3	80.34 ± 0.2	**86.96 ± 0.1**
Gauss. Blur	84.77 ± 0.3	90.65 ± 0.2	87.08 ± 0.4	78.56 ± 0.1	91.01 ± 0.3	84.96 ± 0.2	85.34 ± 0.1	88.46 ± 0.2	**91.03 ± 0.1**
Glass Blur	85.32 ± 0.3	90.92 ± 0.3	83.21 ± 0.1	88.44 ± 0.2	91.19 ± 0.3	85.56 ± 0.1	85.44 ± 0.3	91.00 ± 0.1	**91.35 ± 0.2**
Defocus Blur	83.54 ± 0.2	90.80 ± 0.2	87.23 ± 0.2	80.63 ± 0.4	91.68 ± 0.1	85.45 ± 0.3	85.48 ± 0.2	**93.52 ± 0.2**	91.29 ± 0.3
Motion Blur	81.56 ± 0.1	85.72 ± 0.3	82.58 ± 0.2	83.63 ± 0.4	**88.81 ± 0.3**	85.72 ± 0.2	81.75 ± 0.2	88.51 ± 0.2	86.49 ± 0.1
Zoom Blur	76.80 ± 0.2	76.82 ± 0.2	81.90 ± 0.3	80.25 ± 0.1	87.90 ± 0.4	81.52 ± 0.3	72.28 ± 0.2	**89.02 ± 0.3**	83.84 ± 0.2
Fog	57.87 ± 0.2	54.68 ± 0.3	60.73 ± 0.1	55.27 ± 0.2	74.83 ± 0.3	60.79 ± 0.4	54.55 ± 0.3	76.28 ± 0.2	**86.79 ± 0.2**
Contrast	40.27 ± 0.3	56.61 ± 0.3	51.69 ± 0.4	43.87 ± 0.1	54.16 ± 0.3	53.06 ± 0.4	59.10 ± 0.1	58.72 ± 0.3	**60.04 ± 0.2**
Brightness	00.11 ± 0.3	65.60 ± 0.3	64.86 ± 0.2	**67.40 ± 0.4**	63.60 ± 0.4	65.68 ± 0.1	06.60 ± 0.2	64.09 ± 0.1	64.54 ± 0.3
Saturate	83.22 ± 0.2	91.91 ± 0.1	93.96 ± 0.4	93.93 ± 0.3	**95.75 ± 0.2**	90.44 ± 0.3	84.10 ± 0.3	94.10 ± 0.2	93.00 ± 0.1
JpegComp.	82.95 ± 0.1	83.43 ± 0.1	91.73 ± 0.1	87.85 ± 0.3	85.98 ± 0.2	89.13 ± 0.4	85.85 ± 0.1	89.63 ± 0.3	**92.81 ± 0.2**
Elastic Trans.	83.32 ± 0.1	91.19 ± 0.2	89.20 ± 0.4	91.33 ± 0.4	**94.38 ± 0.3**	86.97 ± 0.5	84.98 ± 0.3	93.58 ± 0.2	92.46 ± 0.1
Avgerage	68.77	78.13	81.14	79.68	84.20	81.05	75.14	85.52	**86.86**
↓	14.49	13.84	11.53	14.30	10.54	11.63	10.03	9.28	**07.91**

↓ denotes that compared with the clean data (original) the average dice descent by.

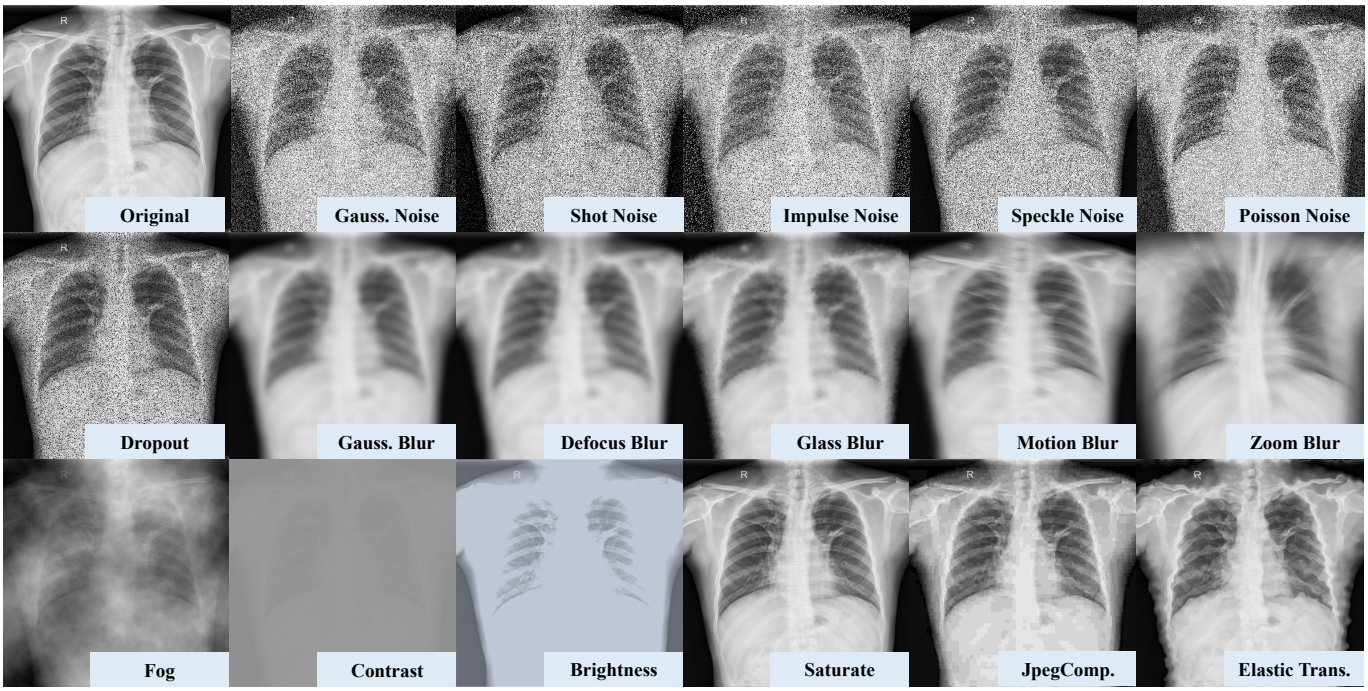

Fig. 1: The visualization of different data corruptions and perturbations. We choose these noises, blur, and digital transformations that are common in real-world applications for verification of the robustness of our model.

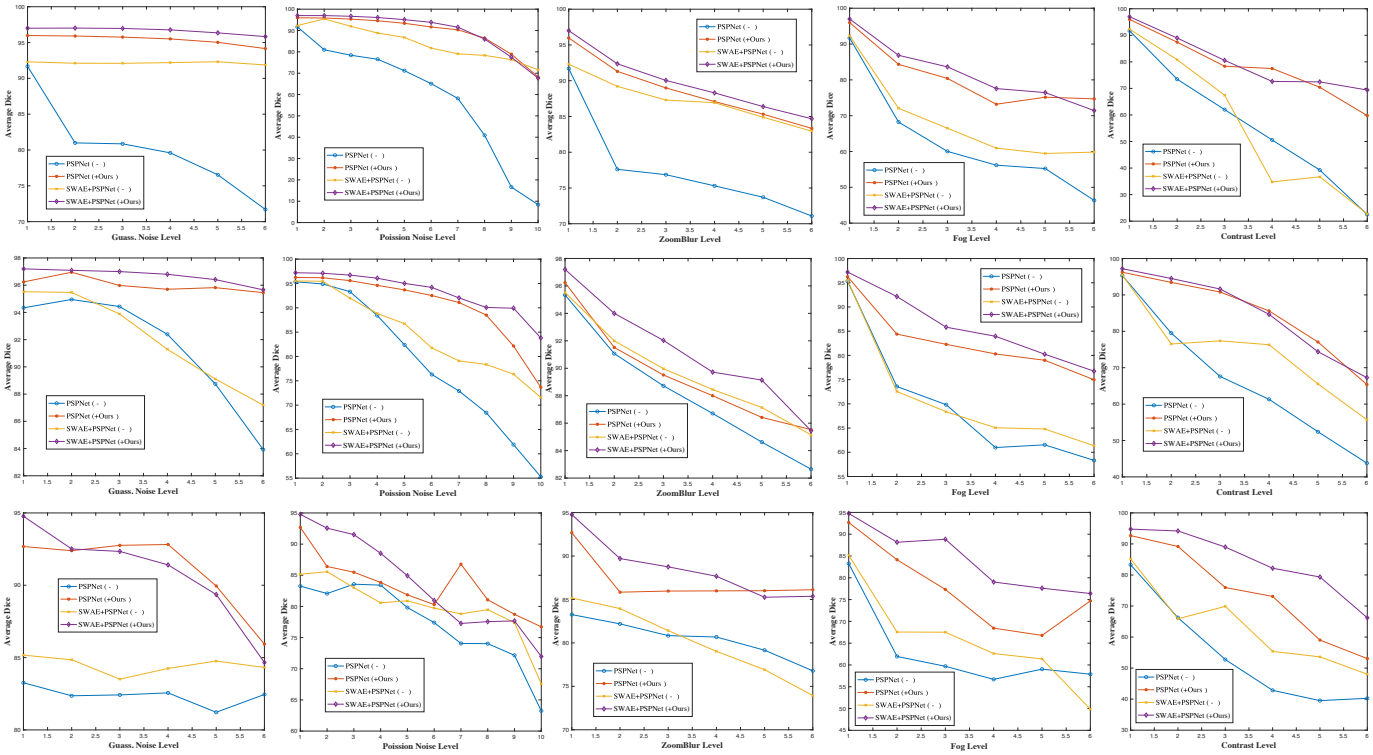

Fig. 2: The robustness performance comparison between our proposed method and baselines on different data corruption. The experiments conducted on the MC dataset are reported in the first row; on the JSRT dataset are reported in the second row; the results on the SH dataset are shown in the third row.

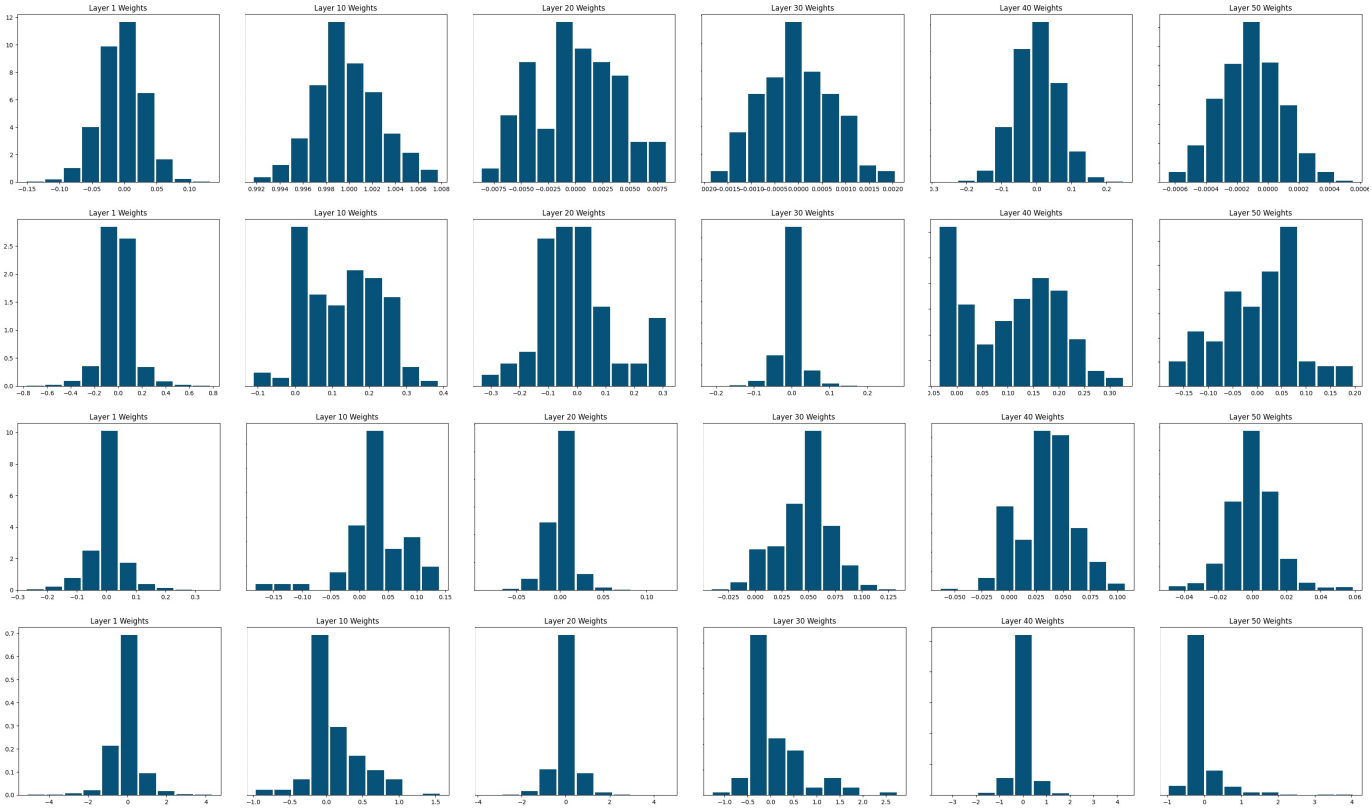

Fig. 3: The visualization of the network's weights distribution. Each column denotes one layer weights. The first row shows weights of ResNet-50 trained from scratch; the second row is with the ImageNet pretrained model; the third one is with the MoCo; the last row denotes utilizing our T-AE.

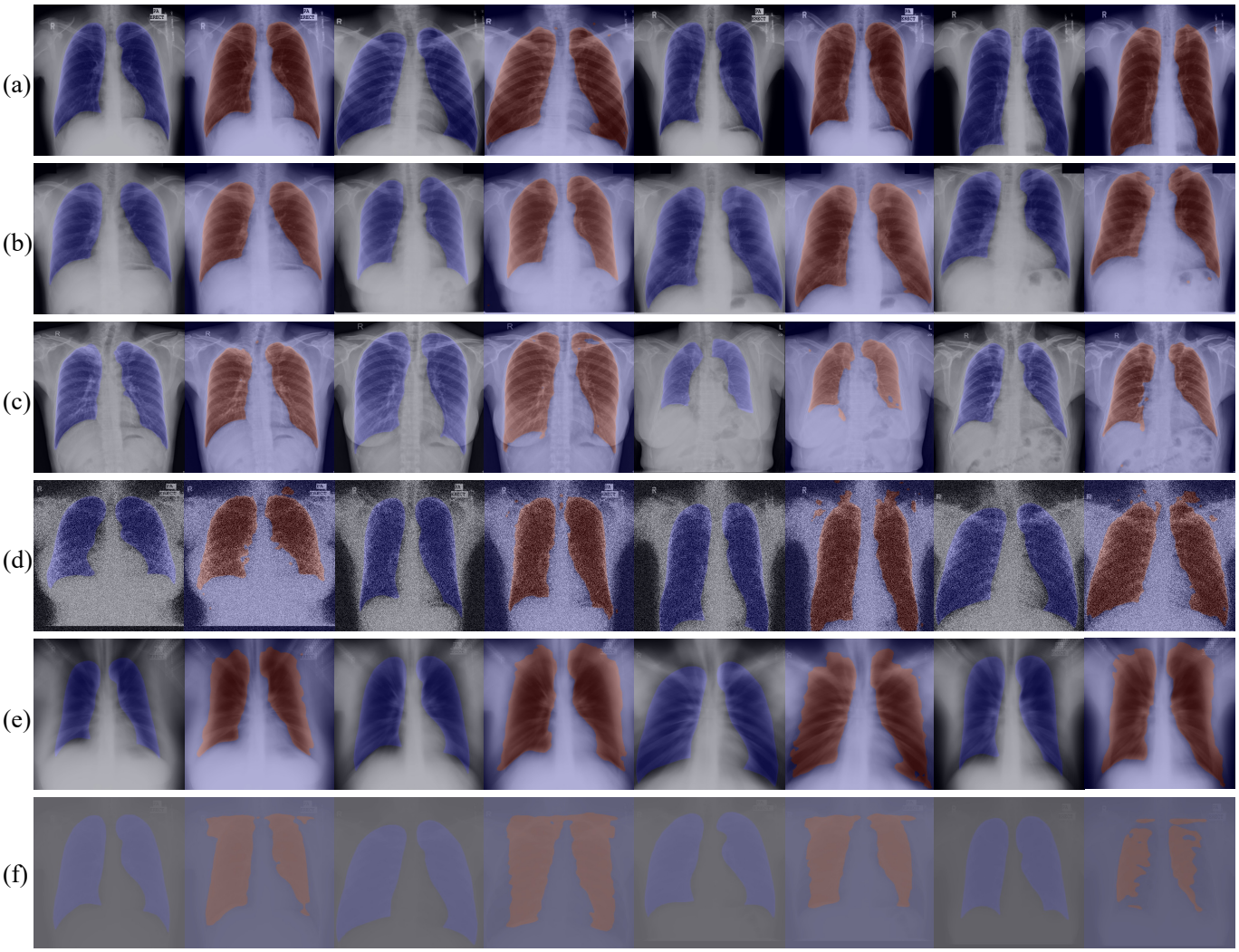

Fig. 4: The visualization of segmentation results. The blue region denotes the mask, and the red area is the predicted result. The model is trained on M dataset, then tests on (a) MC validation dataset; (b) JSRT dataset; (c) SH dataset; (d) Poisson noised MC dataset; (e) zoom blurred MC dataset; (f) contrast transformed MC dataset.