U-Net	labeled/unlabeled CTs	metric	fold 0	fold 1	fold 2	fold 3	fold 4	average
real	101/0	DSC (%)	55.86	52.26	67.34	53.06	59.63	57.63
		NSD (%)	56.87	49.02	68.54	55.02	61.06	58.10
synt	0/116	DSC (%)	61.83	50.38	69.63	57.75	59.46	59.81
		NSD (%)	64.50	47.74	71.48	61.96	60.22	61.28
real & synt	50/52	DSC (%)	56.96	48.77	68.65	54.16	55.76	56.86
		NSD (%)	59.09	43.21	69.44	54.01	54.56	56.06
Swin UNETR-Tiny	labeled/unlabeled CTs	metric	fold 0	fold 1	fold 2	fold 3	fold 4	average
real	101/0	DSC (%)	52.88	49.24	67.94	53.93	55.63	55.92
		NSD (%)	51.34	47.08	71.22	54.56	58.53	56.55
synt	0/116	DSC (%)	55.90	49.63	62.20	52.48	55.30	55.10
		NSD (%)	59.97	46.92	63.23	54.08	53.88	55.61
Swin UNETR-Small	labeled/unlabeled CTs	metric	fold 0	fold 1	fold 2	fold 3	fold 4	average
real	101/0	DSC (%)	60.01	50.56	69.83	52.08	59.98	58.49
		NSD (%)	64.40	48.67	71.20	55.34	59.68	59.86
synt	0/116	DSC (%)	57.16	52.16	63.63	54.79	54.13	56.37
		NSD (%)	63.61	50.04	66.89	57.66	52.98	58.24
Swin UNETR-Base	labeled/unlabeled CTs	metric	fold 0	fold 1	fold 2	fold 3	fold 4	average
real	101/0	DSC (%)[†]	55.35	50.32	64.41	54.17	55.35	55.92
		DSC (%)	59.19	54.04	68.32	52.58	60.97	59.02
		NSD (%)	63.56	52.46	70.06	55.19	62.85	60.82
synt	0/116	DSC (%)	55.26	51.43	64.87	53.34	54.82	55.94
		NSD (%)	62.08	49.87	67.89	57.56	53.61	58.20

[†] The 5-fold cross validation results are provided by Tang *et al*. [55].

Table 5. **Performance on 5-fold cross-validation.** We compare the model (U-Net, Swin-UNETR-Tiny, Small, Base) trained on synthetic tumors with the model trained on real tumors with 5-fold cross-validation. We use Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD) as evaluation metrics to measure tumor segmentation performance. AI models trained solely on synthetic tumors achieve comparable performance to those trained on per-voxel annotation. Furthermore, the U-Net architecture can even exceed the performance of per-voxel annotation. The results indicate that synthetic tumors have the potential to serve as an alternative to real tumors for training AI models. This also signifies a paradigm shift in liver tumor segmentation, transitioning from a label-intensive AI development to a label-free one.

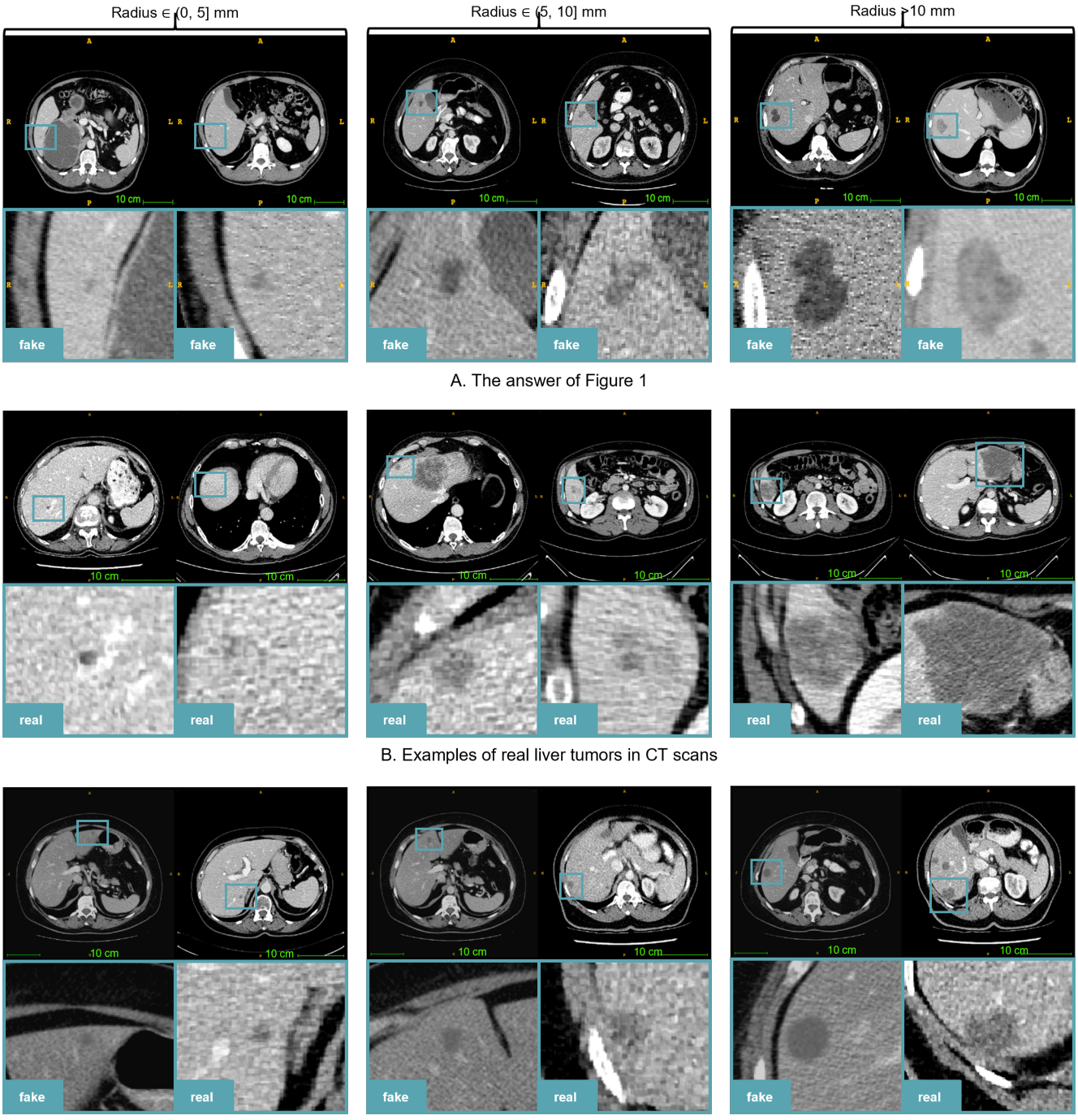

Figure 7. **The answer of Figure 1**. **A.** All the six examples in Figure 1 are synthetic liver tumors generated by our algorithm. **B.** Examples of real liver tumors stratified by tumor size (small, medium, large). **C.** Examples of the Visual Turing Test for clinical validation. These CT scans are sent to medical professionals (format as `nii.gz`). The professionals are asked to mark each CT scan as real, synthetic, or unsure. Based on results in §5.1 and Table 2, the senior professional achieves an accuracy of 26.5% with 1 out of 50 CT scans marked unsure, the junior professional achieves an accuracy of 71.0% with 19 out of 50 marked unsure.

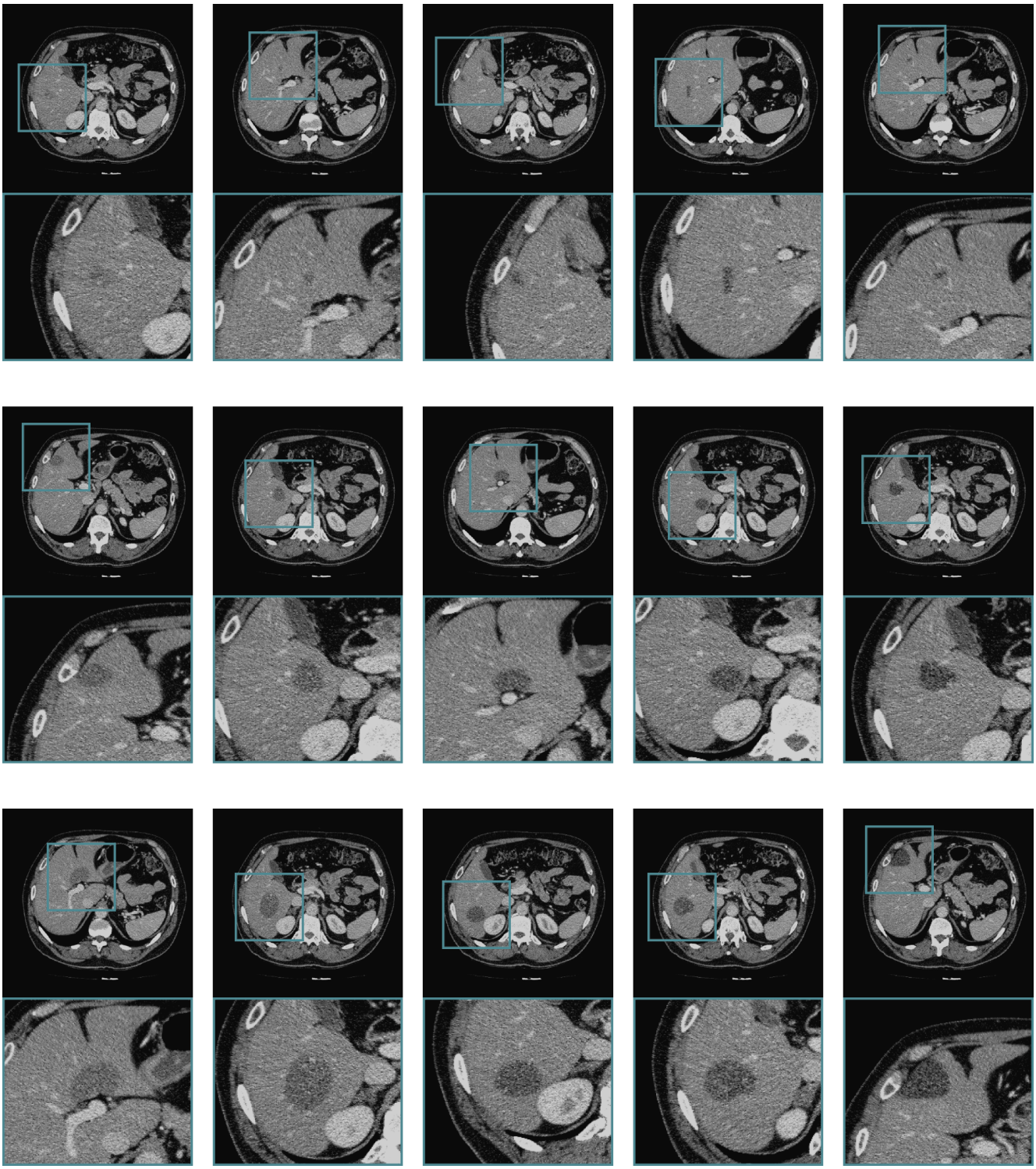

Figure 8. **Visualization of tumor generation: examples.** We have developed a hand-craft strategy to generate synthetic liver tumors. Our synthetic tumors are realistic in shape and texture, which even medical professionals can confuse with real tumors. On the other hand, the generation pipeline is quite flexible, we can control its shape, size, texture, intensity, and location. This figure shows some examples of synthetic tumors generated by our method. The size of the synthetic tumor exhibits an increase from top to bottom, and its intensity becomes darker from left to right.

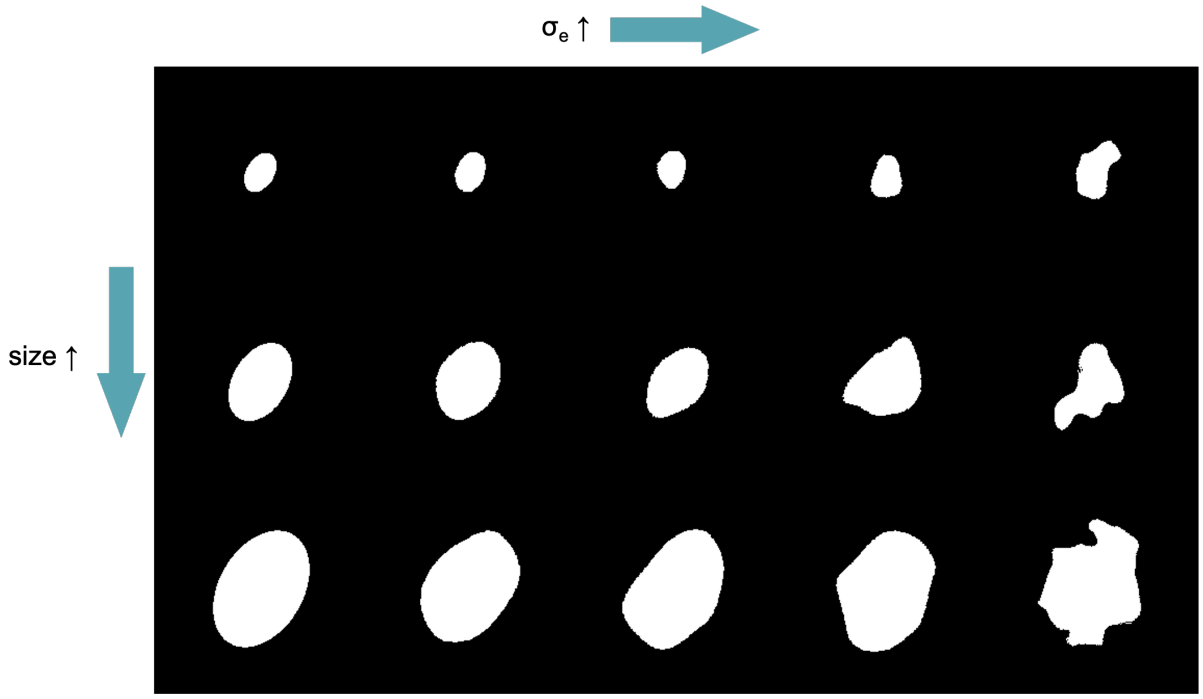

Figure 9. **Visualization of tumor generation: shape.** We show the effect of parameters in "Mask Shape Generation" (Figure 3). The mask shape is controlled by the size r and deformation σ_e. With the increase of r and σ_e, the tumor mask shape becomes larger and more irregular. By choosing appropriate numbers, we are able to simulate real tumor shapes.

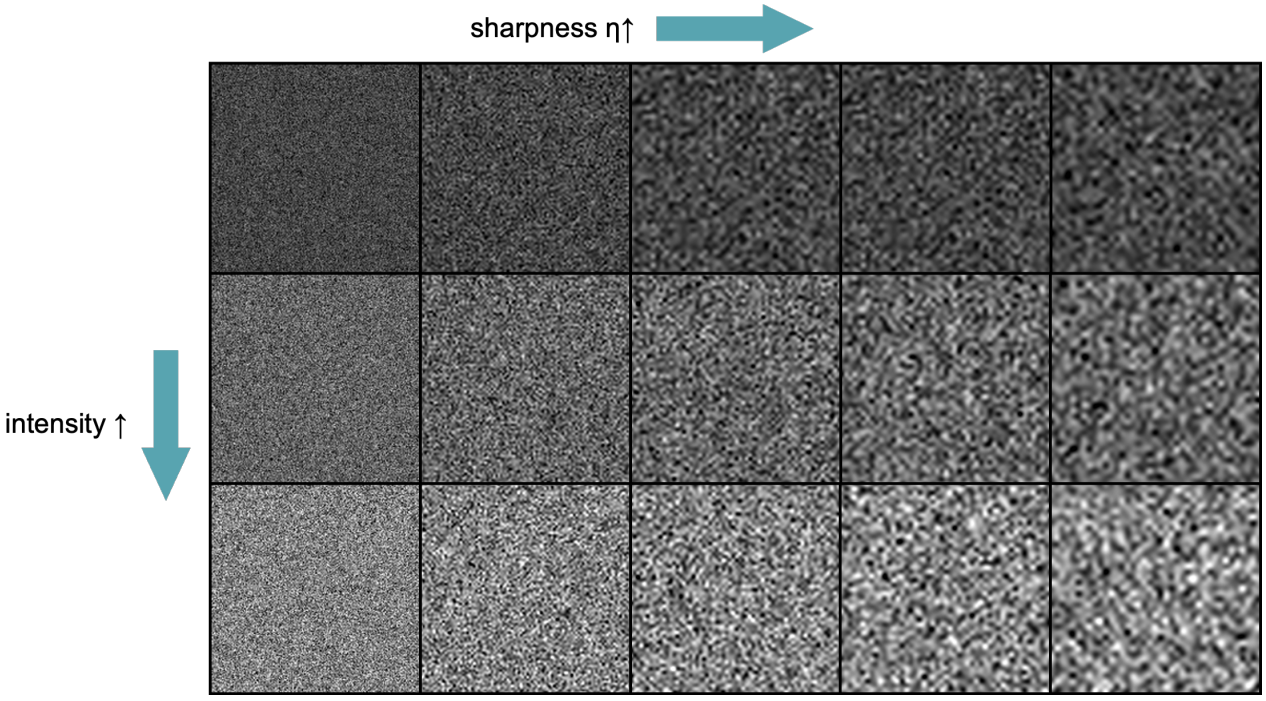

Figure 10. **Visualization of tumor generation: texture.** We show the effect of parameters in "Texture Generation" (Figure 3). The texture of our synthetic tumor is mainly controlled by the intensity μ_t and sharpness η. μ_t represents our synthetic tumor's mean HU value, and η determines how rough the generated texture feels. The hyper-parameters we use to simulate real texture can be found in Table 1.

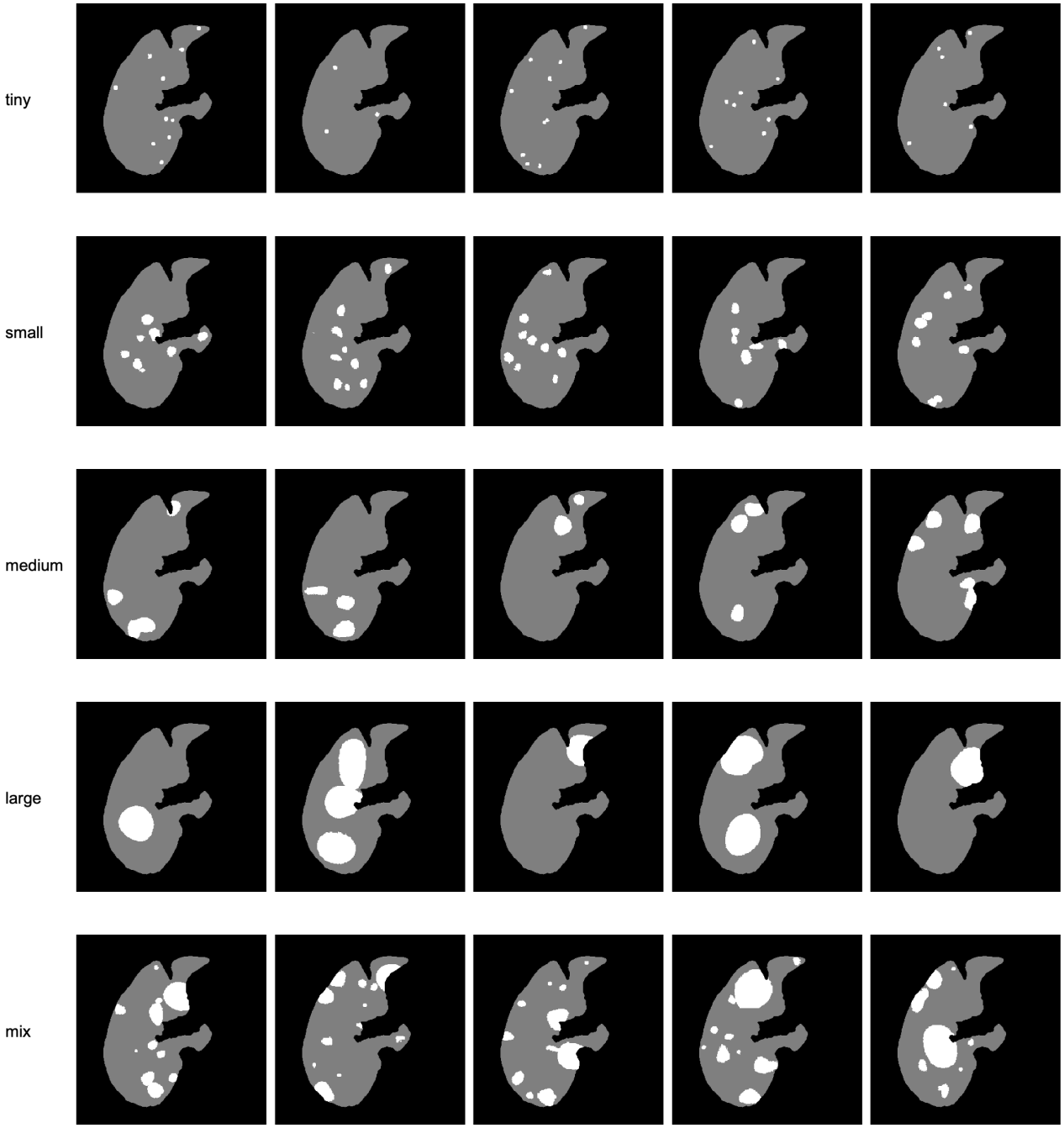

Figure 11. **Visualization of tumors for model training.** During training time, we are able to generate liver tumors on the fly, theoretically creating infinite image-label pairs. We show some visualization examples of "tiny", "small", "medium", "large", and "mix" tumors. The parameters of these tumors are shown in Table 6.

parameter	tiny	small	medium	large	mix
size r	4	8	16	32	/
deformation σ_e	$U\,[0.5, 1]$	$U\,[1, 2]$	$U\,[3, 6]$	$U\,[5, 10]$	/
number N	$F\,[3, 10]$	$F\,[3, 10]$	$F\,[2, 5]$	$F\,[1, 3]$	/

Table 6. **Tumor parameters for model training.** Let $U\,[a, b]$ denotes a uniform distribution, $F\,[a, b]$ denotes a discrete uniform distribution, N denotes synthetic numbers. To train an AI model, we design 5 different types of tumor sizes, tiny, small, medium, large, and mix combine all. The sample probability during training is [0.2, 0.2, 0.2, 0.2, 0.2], respectively.

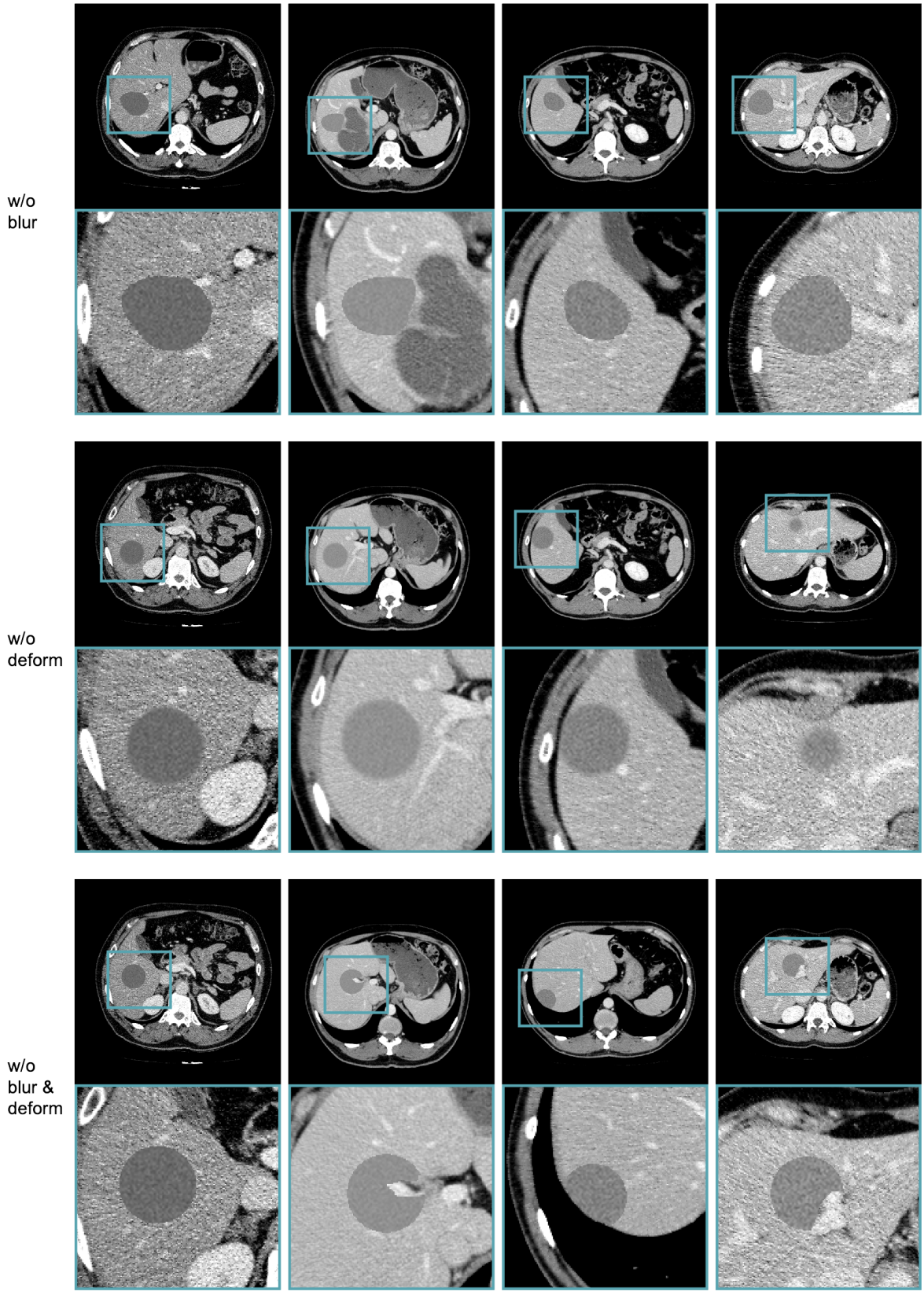

Figure 12. **Visualization of shape ablation.** To show the importance of synthetic shape, we design ablation studies on "Mask Shape Generation" (Figure 3). Without edge blurring and elastic deformation, the edge is sharp and the shape can only be ellipsoid. Therefore, synthetic tumors can be extremely unrealistic.