Learning Segmentation from Radiology Reports

Pedro R. A. S. Bassi[1,2,3], Wenxuan Li[1], Jieneng Chen[1], Zheren Zhu[4,5], Tianyu Lin[1], Sergio Decherchi[3], Andrea Cavalli[2,3,6], Kang Wang[4], Yang Yang[4], Alan L. Yuille[1], and Zongwei Zhou[⋆1]

[1] John Hopkins University
[2] University of Bologna
[3] Italian Institute of Technology
[4] University of California, San Francisco
[5] University of California, Berkeley
[6] École Polytechnique Fédérale de Lausanne

Abstract. Tumor segmentation in CT scans is key for diagnosis, surgery, and prognosis, yet segmentation masks are scarce because their creation requires time and expertise. Public abdominal CT datasets have from dozens to a couple thousand tumor masks, but hospitals have hundreds of thousands of tumor CTs with radiology reports. Thus, leveraging reports to improve segmentation is key for scaling. In this paper, we propose a report-supervision loss (R-Super) that converts radiology reports into voxel-wise supervision for tumor segmentation AI. We created a dataset with 6,718 CT-Report pairs (from the UCSF Hospital), and merged it with public CT-Mask datasets (from AbdomenAtlas 2.0). We used our R-Super to train with these masks and reports, and strongly improved tumor segmentation in internal and external validation—F1 Score increased by up to 16% with respect to training with masks only. By leveraging readily available radiology reports to supplement scarce segmentation masks, R-Super strongly improves AI performance both when very few training masks are available (e.g., 50), and when many masks were available (e.g., 1.7K). Project: https://github.com/MrGiovanni/R-Super

Keywords: Tumor Segmentation · Radiology Reports · Loss Function.

1 Introduction

Tumor segmentation in CT scans is key for diagnosis, surgery, and radiotherapy planning. However, creating voxel-wise segmentation masks is time-consuming and expensive [25,21,8,9] . Thus, most public CT datasets have less than 100 tumor masks, few exceed 1,000, and none exceeds 2,000 masks per tumor type [4,13,1]. Creating segmentation masks is not part of a radiologist's everyday job, but creating radiology reports (texts explaining the findings in a CT) is. Thus, hospitals have hundreds of thousands of CT-Report pairs, and recent public datasets have more than 75K CT-Report pairs (50K chest CT-Report pairs in

⋆ Correspondence to: Zongwei Zhou (zzhou82@jh.edu)

CT-RATE [12], 25K Abdominal CT-Report pairs in Merlin [5]). Furthermore, the information inside reports is valuable for segmentation: tumor count, location (organ / organ sub-segment), and sizes. Radiology reports seem to be a large-scale and valuable source of information, but *can reports improve tumor segmentation?* To answer this, studies extracted classification labels from reports (e.g., tumor present / absent) and used them in multi-task learning for segmentation and classification [27]. Studies also used reports in contrastive language-image pre-training (CLIP) [5]. These methods use reports to optimize auxiliary tasks (classification / CLIP), not to optimize segmentation itself. However, the benefit of auxiliary tasks to segmentation is indirect and sometimes minor (Fig. 4).

We hypothesize that radiology reports can be used to *directly supervise* tumor segmentation. We propose R-Super (**R**eport-**Super**vision), a training paradigm that introduces two novel loss functions—Volume Loss (§2.1) and Ball Loss (§2.2)—to enforce alignment between segmentation outputs and tumor attributes extracted from reports (count, size, and location). This information is parsed by a large language model (LLama 3.1 [10]) using radiologist-designed prompts. R-Super applies to any model architecture and, unlike prior methods that rely on auxiliary tasks, R-Super supervises segmentation directly.

To train R-Super, we built a massive dataset sourced from the University of California San Francisco (UCSF) Hospital and nearby institutions (California, USA). Like most hospitals, UCSF could not provide segmentation masks, only reports. However, UCSF-Train has more tumor CTs (4,967) than any public abdominal CT dataset [16,17,19,20,24]. We trained R-Super using the UCSF-Train reports, in addition to few (50), medium (344) or many (1.7K) tumor segmentation masks (sourced from AbdomenAtlas 2.0 [3]). After training, we conducted both internal validation (UCSF-Test), and external validation on a hospital unseen during training. R-Super strongly improved segmentation and surpassed the 5 state-of-the-art methods in Tab. 1. Our main contributions are:

1. We created loss functions that optimize segmented tumors to be consistent with radiology reports. To the best of our knowledge, our R-Super losses are the first to use radiology reports to directly optimize segmentation in CT.
2. R-Super uses reports to achieve better tumor segmentation with *few* (50), *medium* (344), and *many* (1.7K) segmentation masks in training (up to +16%, +9.2%, and +4.3% F1-Score, respectively, Fig. 4).
3. R-Super improved tumor segmentation both in the hospital that provided the radiology report for training (up to +16% F1-Score, Fig. 4), and on an unseen hospital (up to +9.2% F1-Score, Fig. 4).

2 R-Super

R-Super trains AI to segment tumors that are consistent with the tumor information in radiology reports. **First**, it extracts (and stores) this information—tumor count, locations (organ/organ sub-segment), and diameters—from reports using a zero-shot LLM, Llama 3.1 70B AWQ. It uses radiologist-designed prompts

Table 1. Comparison with related work. We compare R-Super with five state-of-the-art methods. **(1)** Standard segmentation uses only CTs with manual masks. **(2)** Models Genesis applies self-supervision on all CTs but ignores reports. **(3)** CLIP-Like and **(4)** Multi-task learning use reports for auxiliary tasks (e.g., classification), not for direct segmentation. **(5)** Report-guided pseudo-labels reduce false positives but miss false negatives and ignore tumor size. In contrast, R-Super uses detailed report information (tumor count, size, location) to directly supervise segmentation and penalize both false positives and false negatives.

training paradigm	learns from CT w/o mask	learns w/ reports	uses detailed report info.†	reports optimize segmentation directly	reports penalize FP/FN directly
Segmentation [14,11]	✗	✗	✗	✗	✗
Models Gen. [28,29]	✓	✗	✗	✗	✗
Multi-task Learn. [7]	✓	✓	✗	✗	✗
RG pseudo-labels [6]	✓	✓	✗	✓	✗
CLIP-Like [5]	✓	✓	✓	✗	✗
R-Super (ours)	✓	✓	✓	✓	✓

† AI learns from tumor count, sizes, and locations (organ/organ sub-segment) in reports.

and has 96% accuracy extracting tumor presence & location [3]. **Second**, we train the tumor segmenter with the available (possibly few) CT-Mask pairs, using traditional segmentation losses (cross-entropy & DSC). **Third,** we fine-tune it with both CT-Mask pairs and CT-Report pairs. For the CT-Mask pairs, we use the traditional segmentation losses; for CT-Report pairs, we use two novel losses: *Volume Loss* (§2.1) and *Ball Loss* (§2.2). The Volume Loss is applied to an intermediate layer of the segmenter (deep supervision), and it constricts segmented tumors to match the tumor locations and overall tumor volumes extracted from reports. The Ball Loss is stricter, and is applied to the last layer of the segmenter. It optimizes each segmented tumor individually, making them match the tumor count, locations, volumes and diameters in the report. *Overall, R-Super transforms text-wise report supervision into voxel-wise supervision.*

2.1 Volume Loss

The Volume Loss (Fig. 1) enforces consistency between the total tumor volume in the segmentation and in reports. Instead of volumes, reports usually inform tumor diameters, which we use to estimate volumes. Reports provide one to three diameters per tumor. One diameter is common for small, rounder tumors, and two perpendicular diameters are the World Health Organization standard [22]. With one diameter informed (d_1), we estimate tumor volume as a ball volume: $d_1^3\pi/6$. With three diameters (d_1, d_2, d_3), we use the ellipsoid volume: $d_1 d_2 d_z \pi/6$. With two diameters (d_1, d_2), we estimate the third as $d_3 = (d_1 + d_2)/2$, and use the ellipsoid volume. For each organ or organ sub-segment (o), we sum the volumes of all tumors in the report, giving $V_{r,o}$. After estimating $V_{r,o}$ from the report, Eq. 1 estimates $V_{s,o}$, the total volume of all tumors the AI *segmented*

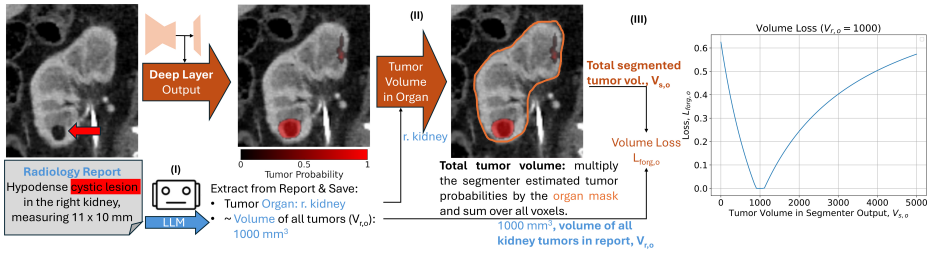

Fig. 1. The Volume Loss (deep supervision) aligns segmented tumors with volumes and locations from reports. It consists of three steps. **(I)** A large language model (LLM) extracts tumor location (organ/sub-segment) and diameters from the report. Using ellipsoid approximations, we estimate the total tumor volume in each organ o as $V_{r,o}$. **(II)** From a deep layer of the segmenter, we compute tumor probabilities via a 1×1×1 convolution and sigmoid/softmax. We sum the probabilities inside the organ (using pre-saved organ masks) to estimate the segmented tumor volume, $V_{s,o}$. **(III)** The Volume Loss penalizes discrepancies between $V_{s,o}$ and $V_{r,o}$. As shown in the right panel (for $V_{r,o} = 1000$), the loss is zero when $V_{s,o}$ is close to $V_{r,o}$, allowing for some tolerance due to uncertainty in report-based volume estimates.

in o. Eq. 1 uses the tumor segmentation, $\boldsymbol{T} = [t_{h,w,l}]$[7]; pre-saved, binary organ segmentation masks, $\boldsymbol{O} = [o_{h,w,l}]$[8]; and the voxel volume v (from CT spacing).

$$V_{s,o} = \sum_{h,w,l}^{H,W,L} t_{h,w,l} o_{h,w,l} v \tag{1}$$

The Volume Loss penalizes the difference between the segmented tumor volume, $V_{s,o}$, and the tumor volume estimated from reports, $V_{r,o}$. To perform this penalization, we tried many losses (e.g., L1 and L2), but Eq. 2 converged better:

$$L'_{\text{forg},o}(V_{s,o}, V_{r,o}) = \frac{|V_{s,o} - V_{r,o}|}{V_{s,o} + V_{r,o} + E} \tag{2}$$

The E term increases stability and ensures the loss minimum is at $V_{s,o} = 0$ when $V_{r,o} = 0$ (we set $E = 500$ mm^3). Importantly, tumor volumes estimated from reports ($V_{r,o}$) are imperfect. Different radiologists may provide slightly different diameters for a tumor, and we use ellipsoid approximations to convert diameters into volumes. Thus, in Eq. 3, we include a tolerance ($0 < \tau < 1$; we set $\tau = 10\%$) in the Volume Loss: if $|V_{s,o} - V_{r,o}| \lesssim \tau V_{r,o}$, the loss and its gradient become 0, not penalizing the segmenter. The Volume Loss also has a

[7] The volume loss is used as deep supervision. Thus, $\boldsymbol{T}$ is created from the output of a deep convolutional layer (e.g., MedFormer decoder layer 2), after a 1x1x1 conv. to reduce channels, sigmoid activation, and linear interpolation to the CT spacing.

[8] Many accurate public organ segmentation models exist [2,25]. We used an nnU-Net [14] trained on AbdomenAtlas 2.0 [3], which we made public. It segments 39 organs / anatomical structures, including the pancreas sub-segments (head, body, and tail) and kidneys. Sub-segment masks improve R-Super, but are not necessary (Tab. 2).

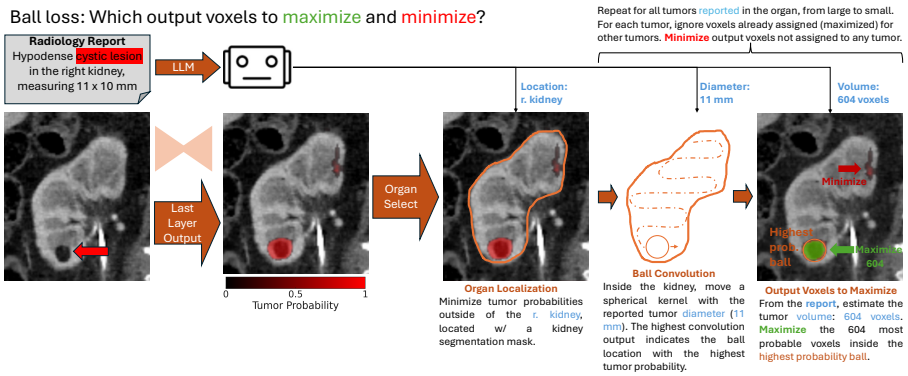

Fig. 2. The Ball Loss enforces segmented tumors to match detailed tumor information from reports—tumor count, locations, diameters, and volumes. It consists of four steps. **(I)** A large language model (LLM) extracts this tumor information from the report. **(II)** A Ball Convolution—a simple convolution with a fixed spherical kernel matching the reported tumor diameter—slides over the segmenter output to find the highest-probability tumor region. **(III)** Within this region, the loss maximizes probabilities in the top-N most probable voxels, where N is the tumor volume estimated from the report (in voxels). These voxels can form any shape that stays within the ball. **(IV)** The process (II–III) is repeated for all reported tumors, largest to smallest. Already-assigned voxels are ignored in later iterations to prevent overlap. Finally, the loss minimizes tumor probabilities in unassigned voxels (background). It focuses more on central tumor voxels and avoids penalizing uncertain tumor borders.

cross-entropy term (Eq. 4) minimizing tumor segmentations in voxels outside the organ / sub-segment containing tumors (o). The final Volume Loss is Eq. 5.

$$L_{\text{forg},o}(V_{s,o}, V_{r,o}) = \max\{L'_{\text{forg},o}(V_{s,o}, V_{r,o}) - L'_{\text{forg},o}((1-\tau)V_{r,o}, V_{r,o}), 0\} \tag{3}$$

$$L_{\text{bkg},o}(\mathbf{T}) = -\frac{1}{H \cdot W \cdot L} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{l=1}^{L} \ln(1 - t_{h,w,l}(1 - o_{h,w,l})) \tag{4}$$

$$L_{\text{vol},o} = L_{\text{forg},o}(V_{s,o}, V_{r,o}) + L_{\text{bkg},o}(\mathbf{T}) \tag{5}$$

2.2 Ball Loss

The Ball Loss is applied to the final segmenter output, and enforces the volume, diameter, location, and count of segmented tumors to match reports. The Volume Loss (deep supervision) does not penalize the count or diameter of segmented tumors, allowing the segmenter to explore different possibilities in its intermediate layers. The Ball Loss makes the segmenter refine its prediction.

Fig. 2 shows the Ball Loss. First, we mask the tumor segmentation output ($\mathbf{T}$) with the corresponding organ's segmentation mask $\mathbf{O}$, giving $\mathbf{T} \otimes \mathbf{O}$. Then, we use Ball Convolutions to locate tumors. This convolution has a non-learnable

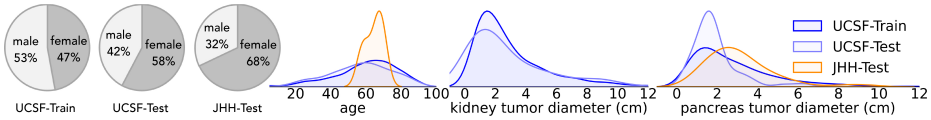

Fig. 3. UCSF-Train has more pancreatic and kidney tumor CTs than any public CT dataset [3]. We show demographic and tumor data for JHH and UCSF train and test sets (unavailable in AbdomenAtlas 2.0). Reports can be used to easily assemble large datasets [18], which R-Super directly uses for segmentation training.

kernel: a ball matching the tumor diameter in the report (the tumor's largest diameter plus a small margin). For now, assume the kernel is binary: 1 inside the ball, and 0 outside. Convolving the ball over the segmentation output sums tumor probabilities (softmax/sigmoid) within each ball position. Where the sum is highest, we infer the most likely location for a tumor with a diameter similar to the ball. We call the ball in this location the "*highest probability ball*". To improve tumor localization, instead of using a binary kernel we weigh the ball kernel higher toward its center, so the convolution responds strongly if the tumor center (where probabilities are usually higher) aligns with the kernel center[9].

First, we use the Ball Convolution to locate the largest tumor the report mentions in the organ o. The convolution provides the tumor's highest probability ball, and we find the top-N most probable tumor voxels inside it. N is how many voxels the tumor should have, according to the tumor volume estimated from report (§2.1). In a blank segmentation mask, we set the top-N voxels to 1. In the segmenter's output, we zero them, to avoid reuse. We repeat the process with the next largest tumor in the report, until all tumors are added to the segmentation mask. The final mask matches the report in tumor count, locations (organs/organ sub-segments), volumes, and diameters. We use these masks as labels for the segmenter, with standard segmentation losses (cross-entropy & DSC). These dynamic masks improve while the segmenter improves. To compensate for uncertainty and inexact diameters in reports, we give the cross-entropy loss higher weights in tumor voxels with higher predicted tumor probability, and the losses do not penalize a small margin at tumor borders.

We train with CT patches. If a patch has only part of an organ with a reported tumor, our losses will tell the segmenter to find this tumor, which can be outside the patch. Thus, each of our training patches fully covers one target organ. If a second organ has a reported tumor and is partially inside the patch, we do not apply losses to this organ. If the report mentions a tumor, the ball loss boosts the region with the highest (even if low) predicted tumor probability. When the AI output is near zero everywhere, the volume loss helps: it discourages all-zero outputs with strong gradients (Fig. 1). We did not apply our losses to organs where reports mentioned tumors but not tumor sizes (they were few, $<11\%$).

[9] The ball is valued 1 at the center, and decays following a 3D Gaussian with std of $0.75\times$ the ball diameter, d_i. The ball is centered, and the kernel is 0 outside it. For input-output alignment, the convolution has stride 1, zero padding and odd kernels.

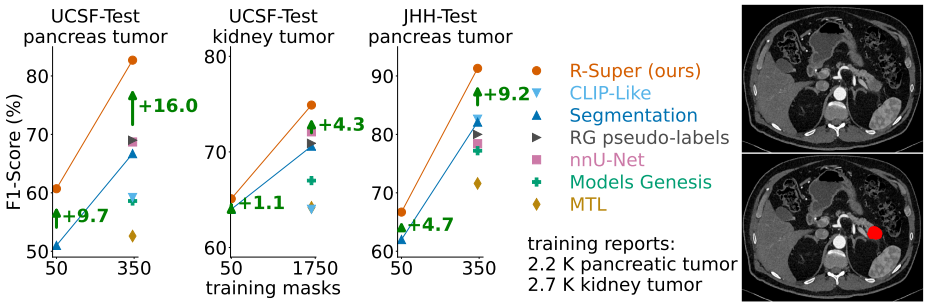

Fig. 4. By learning from reports, R-Super strongly surpassed 6 state-of-the-art methods in internal (UCSF-Test) & external validation (JHH-Test). It surpassed standard segmentation (no report) by up to 16% in tumor detection F1-Score (green arrows). Each plot represents a tumor type and test set; X axes indicate the number of training *masks*. R-Super and report-based AIs were trained with masks and *reports* (2.2K/2.7K pancreas/kidney tumor reports). Test sets are difficult: JHH-Test is from an external hospital and UCSF-Test has many CT resolutions and contrast phases (including plain). On the right, a pancreatic tumor (PDAC) R-Super segmented.

3 Results

Datasets. (I) AbdomenAtlas 2.0 [3] is one of the largest public CT datasets, containing 9,262 CT-Mask pairs from 88 hospitals across 19 countries, including 344 pancreatic and 1,674 kidney tumor CTs. **(II) AbdomenAtlas-Small** is a subset of AbdomenAtlas 2.0, with 50 pancreatic tumor CTs, 50 kidney tumor CTs, and 100 normal CTs. **(III) UCSF-Train** is a large private dataset from UCSF and affiliated institutions (California, USA), with 6,718 CT-Report pairs (no masks): 2,229 with pancreatic tumors, 2,738 with kidney tumors, and 1,751 normals. **(IV) UCSF-Test** is a test set from the same source and distribution as UCSF-Train, consisting of 169 kidney tumor, 139 pancreatic tumor, and 100 normal CTs. **(V) JHH-Test** [26,23] is an external test set from an unseen hospital—Johns Hopkins Hospital (Maryland, USA)—with 50 pancreatic tumor CTs and 50 normals. Tumors are radiologically confirmed focal lesions in I–IV, and pathology-proven PDACs in V. Datasets I–IV have non-contrast and multiple contrast phases; V contains only arterial phase CT. More details are in Fig. 3.

Training R-Super. We trained two versions of R-Super. One using UCSF-Train (CT-Report pairs) & AbdomenAtlas-Small (few CT-Mask pairs), another using UCSF-Train & AbdomenAtlas 2.0 (more CT-Mask pairs). We trained them on CT-Mask pairs, then fine-tuned on CT-Mask & CT-Report pairs. We used standard segmentation losses for CT-Mask, and Volume & Ball losses for CT-Report (weight of 1 for segmentation and 0.1 for our losses; masks and reports balanced w/ data augmentation). We used the MedFormer [11] segmentation architecture and hyper-parameters in R-Super and competing methods—a transformer-CNN hybrid that won benchmarks [2]. The only competing method not using MedFormer was the nnU-Net (ResEncM architecture) [15]. We tested

Table 2. R-Super surpasses state-of-the-art methods in multiple tumor detection and segmentation metrics. For pancreatic tumors, R-Super surpasses all alternative methods in DSC, NSD, F1-Score, AUC. For kidney, R-Super matches segmentation training w/ few masks (50), and surpasses it with many (1.7K). DSC and NSD are not available in UCSF-Test, it does not have masks. Sensitivity (Se) and Specificity (Sp) are for thresholds giving highest F1-Scores. The last 2 table rows are ablations: R-Super with only the volume or the ball loss. R-Super and others were trained with pancreas and kidney tumors. Instead, for a pancreas-only R-Super, Se, Sp, F1, AUC change to 88, 96, 92, 98 in JHH-Test, and 86, 86, 83, 91 in UCSF-Test. Pancreas sub-segment masks improved R-Super; w/o them (full pancreas masks), F1, AUC drop to 88, 84 in JHH and 74, 80 in UCSF. Orange marks significant gains ($p < 0.05$, ablations excluded) in F1 (paired permutation test) and AUC (DeLong's test).

	pancreas tumor												kidney tumor					
	JHH-Test								UCSF-Test				UCSF-Test					
train paradigm	mask	rep.	dsc	nsd	F1	AUC	Se	Sp	F1	AUC	Se	Sp	mask	rep.	F1	AUC	Se	Sp
segmentation [11]	50	0	38	41	62	63	62	62	51	63	47	**77**	50	0	64	**68**	68	**63**
R-Super (ours)	50	2.2K	**49**	**52**	**67**	**75**	**68**	**64**	**61**	**70**	**73**	59	50	2.7K	**65**	66	**73**	55
CLIP-Like [5]	344	2.2K	55	58	83	86	90	70	59	74	59	75	1.7K	2.7K	64	71	75	48
Multi-task l. [7]	344	2.2K	38	43	72	80	78	60	53	62	65	50	1.7K	2.7K	64	71	68	63
RG Pseudo-l. [6]	344	2.2K	56	62	80	79	80	78	69	83	65	86	1.7K	2.7K	71	74	80	60
Models G. [28]	344	0	53	57	77	81	78	74	59	72	66	63	1.7K	0	67	73	77	54
nnU-Net [14]	344	0	53	57	78	75	76	82	69	79	74	74	1.7K	0	72	74	**87**	53
segmentation [11]	344	0	51	59	82	84	78	**88**	67	78	62	85	1.7K	0	71	73	65	68
R-Super (ours)	344	2.2K	**59**	**69**	**91**	**92**	**94**	**88**	**83**	**90**	**83**	**89**	1.7K	2.7K	**75**	**78**	80	**70**
Vol. Loss (ours)	344	2.2K	59	61	94	97	90	98	76	89	80	82	1.7K	2.7K	75	74	86	63
Ball Loss (ours)	344	2.2K	59	59	83	92	78	90	71	82	64	90	1.7K	2.7K	74	75	87	58

on UCSF-Test (internal) and JHH-Test (external), and randomly set 10% of our training set for hold-out validation.

R-Super improves tumor segmentation with few and many training masks. Fig. 4 compares R-Super to standard segmentation (trained w/o reports). By including 2.2K pancreatic and 2.7K kidney tumor CT-Report pairs in training, R-Super strongly surpassed segmentation when the training set had few (50, AbdomenAtlas-Small) training masks (up to +9.7% tumor detection F1-Score), medium (344, AbdomenAtlas 2.0) masks (up to +16% F1-Score), and many (1.7K, AbdomenAtlas 2.0) masks (+4.3% F1-Score). Thus, R-Super can use radiology reports to help segmenting tumor types with scarce segmentation masks, and it can scale up already large tumor segmentation datasets.

R-Super improves tumor segmentation on the hospital that provided training reports and on an unseen hospital. Fig. 4 shows that R-Super strongly surpasses segmentation both on UCSF-Test (same hospital as UCSF-Train), and on JHH-Test, a dataset from a hospital never seen during training. Interestingly, results on JHH-Test surpassed UCSF-Set (Fig. 4), possibly due to finer CT slice thickness (0.5 mm vs. many thickness in UCSF-Test), and arterial contrast in JHH-Test (vs. many contrast phases & non-contrast in UCSF-Set). In both test datasets, R-Super strongly surpasses six state-of-the-art methodologies, including four that also use reports (Tab. 1). R-Super was supe-

rior in F1-Score (Fig. 4), AUC, DSC and NSD (Tab. 2). Thus, its advantages over alternative methods (explained in Tab. 1) translated into better segmentation.

4 Conclusion

R-Super transforms radiology reports into supervision for segmentation, allowing us to scale segmentation AI by leveraging the vast amount of CT–Report pairs in hospitals and public datasets [12,3,5]—when few or many masks are available. It also allows easier merging of data from many hospitals, enhancing data diversity and AI generalization. Here, merging the UCSF-Train CT-Report dataset with AbdomenAtlas 2.0 yielded a training set with 2,573 pancreatic and 4,412 kidney tumor CTs—way more than any public set. Yet, a single hospital can have orders of magnitude more reports. Therefore, the exceptional R-Super results in this study (e.g., +16% F1-Score in Fig. 4) are an initial sample—and an unveiling—of the transformative potential of report-driven segmentation AI.

Acknowledgments. This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research, the Patrick J. McGovern Foundation Award, and the National Institutes of Health (NIH) under Award Number R01EB037669. We would like to thank the Johns Hopkins Research IT team in IT@JH for their support and infrastructure resources where some of these analyses were conducted; especially DISCOVERY HPC. We thank funding from CBN, IIT (73010, Arnesano, LE, IT).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., van Ginneken, B., et al.: The medical segmentation decathlon. arXiv preprint arXiv:2106.05735 (2021)
2. Bassi, P.R., Li, W., Tang, Y., Isensee, F., et al.: Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation? Conference on Neural Information Processing Systems (2024), https://github.com/MrGiovanni/Touchstone
3. Bassi, P.R., Yavuz, M.C., Wang, K., Chen, X., Li, W., Decherchi, S., Cavalli, A., Yang, Y., Yuille, A., Zhou, Z.: Radgpt: Constructing 3d image-text tumor datasets. arXiv preprint arXiv:2501.04678 (2025), https://github.com/MrGiovanni/RadGPT
4. Bilic, P., Christ, P.F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., Fu, C.W., Han, X., Heng, P.A., Hesser, J., et al.: The liver tumor segmentation benchmark (lits). arXiv preprint arXiv:1901.04056 (2019)
5. Blankemeier, L., Cohen, J.P., Kumar, A., Van Veen, D., Gardezi, S.J.S., Paschali, M., Chen, Z., Delbrouck, J.B., Reis, E., Truyts, C., et al.: Merlin: A vision language foundation model for 3d computed tomography. Research Square pp. rs–3 (2024)
6. Bosma, J.S., Saha, A., Hosseinzadeh, M., Slootweg, I., de Rooij, M., Huisman, H.: Semisupervised learning with report-guided pseudo labels for deep learning–based prostate cancer detection using biparametric mri. Radiology: Artificial Intelligence **5**(5), e230031 (2023)

7. Chen, E.Z., Dong, X., Li, X., Jiang, H., Rong, R., Wu, J.: Lesion attributes segmentation for melanoma detection with multi-task u-net. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). pp. 485–488. IEEE (2019)
8. Chou, Y.C., Li, B., Fan, D.P., Yuille, A., Zhou, Z.: Acquiring weak annotations for tumor localization in temporal and volumetric data. Machine Intelligence Research pp. 1–13 (2024), https://github.com/johnson111788/Drag-Drop
9. Chou, Y.C., Zhou, Z., Yuille, A.: Embracing massive medical data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 24–35. Springer (2024), https://github.com/MrGiovanni/OnlineLearning
10. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
11. Gao, Y., Zhou, M., Liu, D., Yan, Z., Zhang, S., Metaxas, D.N.: A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark. arXiv preprint arXiv:2203.00131 (2022)
12. Hamamci, I.E., Er, S., Menze, B.: Ct2rep: Automated radiology report generation for 3d medical imaging (2024), https://arxiv.org/abs/2403.06801
13. Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al.: The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445 (2019)
14. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
15. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. arXiv preprint arXiv:2404.09556 (2024)
16. Jaus, A., Seibold, C., Hermann, K., Walter, A., Giske, K., Haubold, J., Kleesiek, J., Stiefelhagen, R.: Towards unifying anatomy segmentation: Automated generation of a full-body ct dataset via knowledge aggregation and anatomical guidelines. arXiv preprint arXiv:2307.13375 (2023)
17. Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. arXiv preprint arXiv:2206.08023 (2022)
18. Li, W., Bassi, P.R., Lin, T., Chou, Y.C., Zhou, X., Tang, Y., Isensee, F., Wang, K., Chen, Q., Xu, X., et al.: Scalemai: Accelerating the development of trusted datasets and ai models. arXiv preprint arXiv:2501.03410 (2025), https://github.com/MrGiovanni/ScaleMAI
19. Li, W., Qu, C., Chen, X., Bassi, P.R., Shi, Y., Lai, Y., Yu, Q., Xue, H., Chen, Y., Lin, X., et al.: Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. Medical Image Analysis p. 103285 (2024), https://github.com/MrGiovanni/AbdomenAtlas
20. Li, W., Yuille, A., Zhou, Z.: How well do supervised models transfer to 3d image segmentation? In: International Conference on Learning Representations (2024), https://github.com/MrGiovanni/SuPreM
21. Ma, J., Zhang, Y., Gu, S., Ge, C., Wang, E., Zhou, Q., Huang, Z., Lyu, P., He, J., Wang, B.: Automatic organ and pan-cancer segmentation in abdomen ct: the flare 2023 challenge (2024), https://arxiv.org/abs/2408.12534
22. Miller, A., Hoogstraten, B., Staquet, M., Winkler, A.: Reporting results of cancer treatment. Cancer **47**(1), 207–214 (1981)

23. Park, S., Chu, L., Fishman, E., Yuille, A., Vogelstein, B., Kinzler, K., Horton, K., Hruban, R., Zinreich, E., Fouladi, D.F., et al.: Annotated normal ct data of the abdomen for deep learning: Challenges and strategies for implementation. Diagnostic and interventional imaging **101**(1), 35–44 (2020)
24. Qu, C., Zhang, T., Qiao, H., Liu, J., Tang, Y., Yuille, A., Zhou, Z.: Abdomenatlas-8k: Annotating 8,000 abdominal ct volumes for multi-organ segmentation in three weeks. In: Conference on Neural Information Processing Systems. vol. 21 (2023), https://github.com/MrGiovanni/AbdomenAtlas
25. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5) (2023)
26. Xia, Y., Yu, Q., Chu, L., Kawamoto, S., Park, S., Liu, F., Chen, J., Zhu, Z., Li, B., Zhou, Z., et al.: The felix project: Deep networks to detect pancreatic neoplasms. medRxiv (2022)
27. Zhang, Y., Li, H., Du, J., Qin, J., Wang, T., Chen, Y., Liu, B., Gao, W., Ma, G., Lei, B.: 3d multi-attention guided multi-task learning network for automatic gastric tumor segmentation and lymph node classification. IEEE transactions on medical imaging **40**(6), 1618–1631 (2021)
28. Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.: Models genesis. Medical Image Analysis **67**, 101840 (2021), https://github.com/MrGiovanni/ModelsGenesis
29. Zhou, Z., Sodha, V., Siddiquee, M.M.R., Feng, R., Tajbakhsh, N., Gotway, M.B., Liang, J.: Models genesis: Generic autodidactic models for 3d medical image analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 384–393. Springer (2019), https://github.com/MrGiovanni/ModelsGenesis