nnQC: a self-adapting framework for quality control in organ segmentation.

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Medical image segmentation, powered by deep learning, has revolutionized automated analysis pipelines for large-scale population studies. However, state-of-the-art methods are prone to hallucinations that lead to anatomically implausible segmentations. With manual correction impractical at scale, automated quality control (QC) techniques have emerged to address the challenge. While promising, existing QC methods are designed for handling a single organ, which restricts their generalizability across different applications. To overcome this limitation, we propose *no-new Quality Control (nnQC)*, a robust QC framework based on a diffusion-generative paradigm that self-adapts to any input organ dataset. Central to nnQC is a novel *Team of Experts (ToE)* architecture, where two specialized *experts* independently process an image and its predicted segmentation, generating a pair of independent embeddings, or *opinions*. A weighted conditional module combines the opinions to guide the sampling within a diffusion process, enabling the accurate generation of a spatially-aware pseudo-ground truth (pGT) used to predict QC scores. We evaluated nnQC on six organs using publicly available datasets. By adapting the network through extracted dataset information, or *fingerprints*, and leveraging the proposed ToE framework, our results demonstrate that nnQC consistently outperforms state-of-the-art methods across all experiments, including cases where segmentation masks are highly degraded or completely missing – confirming it as a versatile and off-the-shelf QC solution across different organs.

Keywords: Quality control · Medical image segmentation · Self-adapting.

1 Introduction

Advances in deep learning (DL) have demonstrated unprecedented capabilities in automating and expediting medical image segmentation [13,17,29]. Despite their high accuracy, DL techniques can still predict anatomically implausible segmentations [3]. Thus, their integration in real-world applications requires visual quality control (QC), which involves inspecting each segmented image for spurious results, followed by correcting or discarding them manually, tasks that become unfeasible at scale [22]. Automated quality control (QC) techniques have emerged as a mechanism to bypass exhaustive visual QC of predicted segmentations [8]. Despite the promise of automated QC, existing methods remain metric-

and organ-specific, making their seamless use across applications difficult [29]. Enabling large-scale population studies requires robust and self-adapting QC frameworks capable of jointly working with state-of-the-art segmentation methods [13] and able to assesssegmentations of varying quality and degradation from different organs.

Related works. Numerous QC frameworks have been proposed, focusing on organ-specific segmentation QC [1,8,9,27,15,4]. Existing QC methods can be embedded in the segmentation method, allowing to self-evaluate the predicted segmentation [14]; semi-detached, where the QC module is separate but tailored for a specific family of segmentation methods [1]; or fully separated. As both embedded and semi-detached QC techniques are inherently tied to the segmentation model they are designed for, this limits their applicability to other segmentation frameworks. Independent QC modules, instead, by being model-agnostic offer more flexibility. Among those, a significant part of the literature focuses on metric-specific approaches, i.e., models trained to predict either qualitative scores (e.g., good/bad) within classification frameworks [16] or quantitative scores (e.g., Dice Score) through regression-based modeling [8,21]. In practice however, these methods are limited by their need for annotations for training [21], or their inability to handle unbounded metrics, such as the Hausdorff Distance. In constrast, reconstruction-based QC approaches [26,9,27] circumvent these limitations by generating a pseudo-ground truth (pGT) mask associated with an image and its corresponding predicted segmentation – that can be used to estimate quality scores of the predicted segmentation. However, as these models typically learn image atlases [26] or manifolds [9,27], they are limited by their reliance on distance metrics to identify the point within the normative model that is closest to a predicted segmentation. When a predicted segmentation is particularly poor, it will lie far from the high-quality points in the normative model, thus, generating pGTs that break the assumption of similarity with the real ground truth and rendering the estimated quality scores unreliable.

Contributions. We introduce *no-new Quality Control* (nnQC), a self-adapting, metric- and model-agnostic QC framework based on latent diffusion models (LDMs) that is robust to varying segmentation qualities and adaptable across organs, datasets, and imaging modalities. The main contributions can be summarized as follows. **(1) Latent Space Sampling via Team of Experts:** we formulate a novel sampling strategy to circumvent the limitations of reconstruction-based QC techniques. The novel strategy dynamically balances two sets of independent (*opinions*), obtained from two independent (*experts*) separately encoding information from the image-segmentation pair, to form a conditional vector to guide the latent vector sampling process of a latent diffusion model (LDM). **(2) Self-Adaptation via Fingerprints:** inspired by nnUNet [13], nnQC extracts dataset-specific attributes, or *fingerprints*, to ensure seamlessly self-adaptation across different organs, datasets, and imaging modalities. **(3) Open-Source Codebase for Open-Science:** to ensure reproducibility, our code and pre-trained weights will be released at github/link.

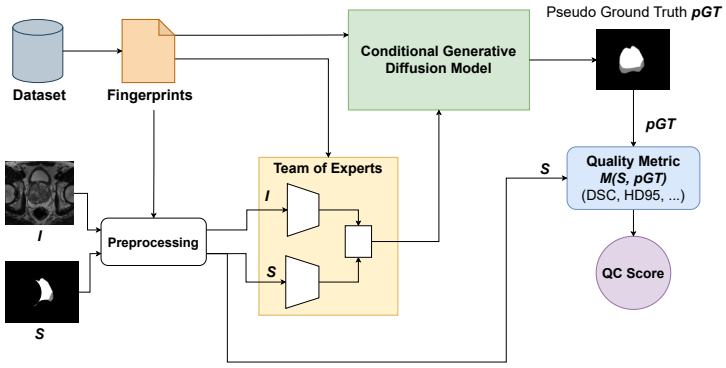

Fig. 1. Overview of the nnQC framework. Given an image-segmentation pair (I, S), a preprocessing module leverages dataset-specific *fingerprints* to normalize the inputs. The *Team of Experts* (ToE) extracts opinions used to condition a latent diffusion model generating a pseudo-ground truth (pGT). The quality score of S is obtained by computing a metric M between S and pGT.

2 Methods

Given an image I and its associated segmentation S generated by an arbitrary segmentation model, we aim to perform segmentation QC by generating a pseudo-ground truth segmentation, pGT_I^S that approximates the real but unknown ground truth segmentation, GT_I, of I. The pseudo-ground truth then serves as a reference for computing the quality score of S using an arbitrary quality metric $M(S, pGT_I^S)$, such that $M(S, pGT_I^S) \simeq M(S, GT_I)$.

In this work, we approach QC by learning a manifold of high-quality segmentations. The learning of nnQC is split into two distinct stages, as in [7]. In a first stage, we train a Team of Experts (ToE) module (Section 2.1) to produce 2 independent embeddings, or *opinions*, used to guide the sampling process of a latent diffusion model (LDM) operating in a latent space derived by a variational autoencoder (VAE) (Section 2.2). In the second stage, the LDM is trained, using the thus pre-trained ToE and VAE, to generate spatially-aware pseudo-ground truths (pGT_I^S) that closely approximate the real ground truth (Section 2.3). Finally, a chosen metric M is computed between S and pGT to obtain the quality score of S. Our framework utilizes dataset-specific attributes, termed *fingerprints* (Section 2.4) to ensure adaptability to different organs, datasets, and imaging modalities. Figure 1 presents an overview of the proposed nnQC framework.

2.1 Team of Experts: Dual Embeddings for QC Conditioning

We enforce nnQC to sample from a good-quality manifold by conditioning an LDM [23] on two complementary feature sets extracted from segmentation maps and images to guide the sampling process for generating pGTs.. Each feature set, or *opinion*, is derived from a specialized feature extractor, or *expert*.

Expert E_1: Processing Segmentations. E_1 is the encoder of a Convolutional Autoencoder (AE) adapted from [9], trained for anomaly detection to reconstruct defect-free segmentations. It minimizes $\mathcal{L}_{AE_1} = \mathcal{L}_{Dice}(S, \hat{S}) + \mathcal{L}_{MSE}(S, \hat{S})$, where $\hat{S}$ is the reconstructed segmentation. $\mathcal{L}_{Dice}$ and $\mathcal{L}_{MSE}$ ensure high-quality reconstructions by preserving spatial relationships even when minor defects are present in the input segmentation.

Expert E_2: Extracting Image-Based Features. E_2 processes the input image to capture essential structural and textural details that complement segmentation features. Using a denoising AE adapted from [2], E_2 minimizes the combined loss $\mathcal{L}_{AE_2} = \mathcal{L}_{SSIM}(I, \hat{I}) + \mathcal{L}_{MSE}(I, \hat{I})$, where $\hat{I}$ is the reconstructed image and $\epsilon \sim \mathcal{N}(0, 1)$ is Gaussian noise added to I. The Structural Similarity Index Measure ($\mathcal{L}_{SSIM}$) captures structural information, while $\mathcal{L}_{MSE}$ ensures accurate pixel-level reconstruction.

Pay Attention to Each Opinion. To dynamically balance the two experts' opinions, we use a Cross-Attention mechanism [6,20]. Each opinion o_1 from E_1 and o_2 from E_2 is first projected using linear layers F_Q, F_K, F_V to produce the query $Q = F_Q(o_1)$, key $K = F_K(o_2)$, and value $V = F_V(o_2)$. The final Cross-Attention vector is:

$$c = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where d_k is the dimensionality of the keys. This mechanism acts as a *dynamic switch*, assigning appropriate importance to each opinion and producing a unified conditioning vector c, a core component of nnQC for handling various quality degradation scenarios.

2.2 Spatial VAE for Good Quality Segmentations

To learn how to express the input samples in a meaningful latent representation, we couple a 2D spatial VAE-GAN [10] with the diffusion model. Following the training strategy outlined in [7], the spatial VAE learns to compress high-quality ground truth, one-hot-encoded binary masks into a latent space $z \in \mathbb{R}^{3\times64\times64}$. The training objective of the spatial VAE is to minimize the following loss:

$$\begin{aligned}\mathcal{L}_{VAE} = \lambda_{KLD}\mathcal{L}_{KLD}\big(VAE_E(S) \parallel \mathcal{N}(0,1)\big) + \\ + \lambda_{perc}\mathcal{L}_{perc}\big(S, \hat{S}\big) \\ + \lambda_{adv}\mathcal{L}_{adv}\big(D(S), D(\hat{S})\big) \\ + \lambda_{Dice}\mathcal{L}_{Dice}\big(S, \hat{S}\big)\end{aligned}$$

where S is the input segmentation map, $\hat{S}$ is the reconstructed segmentation, $\mathcal{L}_{KLD}$ is the Kullback-Leibler divergence loss that forces the latent space $VAE(S)$ to be normally distributed, $\mathcal{L}_{perc}$ denotes a perceptual loss [28], $\mathcal{L}_{Dice}$ represents the generalized Dice Loss, and $\mathcal{L}_{adv}$ is a patch-GAN adversarial loss [10] obtained by forwarding synthetic and real segmentations through a patch-GAN

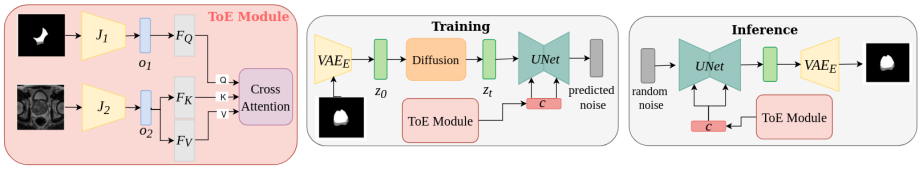

Fig. 2. (Left) The ToE module. (Center) The internal UNet learns to reconstruct noise guided by expert opinions, with conditioning c derived from learnable projections F_Q, F_K, and F_V in the Cross-Attention mechanism. Noise is progressively added to the latent space z_0 to generate z_t. (Right) During inference, a condition c is generated from sampled Gaussian noise to guide the UNet's Cross-Attention module to produce the latent representation z_0. This is decoded by VAE_D to produce the pseudo-ground truth (pGT).

discriminator D [7]. We choose $\mathcal{L}_{Dice}$ as it allows the VAE to learn the spatial relationships among different classes in the input segmentation [9]; $\mathcal{L}_{perc}$ and $\mathcal{L}_{adv}$ are also included due to their proven effectiveness in improving reconstruction quality [7,23]. Each λ_i for the $i-th$ loss $\mathcal{L}_i$ serves as the weight of each loss contribution in the final $\mathcal{L}_{VAE}$ expression.

2.3 Latent Diffusion Models for Segmentation map Generation

For the LDM network, we use a UNet architecture [23,24] as the network that learns the denoising process. In the forward diffusion process, a timestep $t \in [0, T]$ and Gaussian noise $\epsilon \in \mathbb{R}^{3\times64\times64}$ are sampled to corrupt the initial latent representation z_t generated by VAE_E. The UNet is then trained using the denoising loss between ϵ and the predicted noise ϵ_θ, minimizing $\mathcal{L}_{LDM} = ||\epsilon - \epsilon_\theta(z_t; c)||_2^2$, as in [11]. To ensure robustness in handling different corrupted scenario we randomly pass to E_1 mildly to highly corrupted segmentations. We use different morphological manipulations to address that: for each label in the segmentation, we firstly randomly select the percentage of the area to corrupt, then we apply one among erosion or randomly generated holes as the corruption to apply on the mask's portion. At inference, a randomly sampled Gaussian noise is processed by the denoising Unet guided by the dual-condition obtained from the two previously extracted *opinions*, to produce the latent representation z_0. Finally, the predicted and denoised latent sample is given to the VAE_D to create the final pGT_I^S. Figure 2 illustrates the training and inference stages of the LDM.

2.4 Fingerprints for Self-Adaptable QC

Inspired by the nnUNet [13], we use fingerprints to enable our framework to self-adapt to various data types and conditions. We define the *fingerprints* as a set of key characteristics that describe the input dataset: the median voxel spacing of subject volumes, the median size of foreground regions, image orientation,

intensity ranges specific to each modality, and the number of unique segmentation classes. We preprocess the image using median voxel spacing and the median cropped volume size to rescale the volumes accordingly. This is followed by image contrast scaling based on extracted 0.5 and 99.5 percentiles of intensity values within the foreground regions [13] to ensure modality-specific preprocessing. The rescaled images are then aligned to a predefined orientation, producing uniform samples for various training stages. This preprocessing pipeline, inverted during post-processing, guarantees normalized input images for E_2's encoding and standardizes the resolution of the $(I,\ S)$ pairs to 256×256, maintaining a standard foreground area across slices. Unlike the intensive fingerprint-based adaptation process in nnUNet [13], we leverage the intrinsic adaptability of LDMs to operate within a predefined image space [23]. We fix the latent space dimensionality for each latent vector z_t as $z_t \in \mathbf{R}^{3\times64\times64}$ and the embedded dimensions of the *opinions* from E_1 and E_2 as $E_i \in \mathbf{R}^{100\times4\times4}$. Furthermore, we apply minor self-adjustments to the network's first and last layers (E_1, E_2, and the VAE) based on the number of segmentation labels in the dataset.

3 Experiments and Results

We demonstrate the performances of nnQC across various metrics, organs, and imaging modalities. We evaluate nnQC and several competitor models on segmentations generated by nnUNet [13], which are manually degraded to simulate a wide range of quality levels. Degradation is achieved by introducing blank holes into each class of the input masks, applying iterative erosion, or randomly overwriting multi-class segmentations with a single class.

Evaluation Metrics. The performances of nnQC and other competing models are measured using the Pearson correlation (r) between the predicted pseudo-quality scores and real quality scores (i.e., the ones obtained using the pGT and if GT had been available, respectively). We use the Dice Score Coefficient (DSC) and the 95% Hausdorff Distance (HD95) as quality metrics.

Datasets. To assess generalization across varying organ sizes, experiments are conducted on 4 datasets covering five organ types and two imaging modalities (MRI and CT), for a total of 2410 3D volumes. These are: **(1) Prostate Data** from the PI-CAI challenge [25], using 700 bpMRI cases from the PI-CAI data for training and 328 from the PROSTATE-X subset for testing. Experiments are conducted on both full-gland segmentations (1 class) and zonal segmentations (2 classes), specifically targeting the peripheral zone (PZ) and transition zone (TZ). **(2) Cardiac Data** from the M&M 1 and 2 public challenges [5,19], representing a total of 805 MRI scans from six centers, with segmentations for the Left Ventricle (LV), Myocardium (MYO), and Right Ventricle (RV). The models are trained on one dataset and tested on the other. **(3) Spleen and Liver Data** from the FLARE21 public dataset [18], consisting of 511 CT cases from 11 centers. **(4) Brain Data** from the SynthStrip dataset (IXI and Infant subsets) [12], using 50 T1-weighted MRI cases from IXI for training and 16 T1-weighted cases from the Infant subset for testing.

Table 1. Predicted scores/real scores correlation across organs and models

Organ	Class	Score Correlation r							
		Galati et al. [9]		Fournel et al. [8]		Wang et al. [27]		nnQC	
		DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95
Heart	LV	0.75	0.62	0.21	N/A	0.90	0.91	**0.93**	**0.89**
	MYO	-0.06	0.04	0.51	N/A	0.91	0.91	**0.92**	**0.88**
	RV	0.66	0.37	0.58	N/A	0.75	0.71	**0.92**	**0.75**
Prostate	PZ	0.71	0.15	0.33	N/A	0.55	0.51	**0.81**	**0.87**
	TZ	0.26	0.05	0.44	N/A	0.65	0.67	**0.96**	**0.72**
Prostate	Whole	0.26	0.32	0.66	N/A	0.64	0.57	**0.97**	**0.88**
Spleen	Whole	-0.36	-0.2	0.07	N/A	0.71	0.53	**0.95**	**0.80**
Liver	Whole	0.0	0.48	0.58	N/A	0.72	0.57	**0.90**	**0.87**
Brain	Whole	-0.48	-0.13	0.19	N/A	0.42	0.68	**0.83**	**0.96**

Table 2. Accuracy in % for nnQC across different organs in detecting anomalous slices given a pre-defined set of thresholds.

DSC Threshold	Heart	Prostate	Spleen	Liver	Brain
0.6	100	100	99	96	89
0.7	95	98	86	85	75
0.9	82	93	83	83	71

Benchmarks. We consider three baselines: **(1) Fournel et al.** [8] a regression-based model consisting of a ResNet trained as a multi-channel metric regressor, fed with the channel-wise concatenation of image and input segmentation; **(2) Galati et al.** [9] a deterministic reconstructor based on a Convolutional Autoencoder, which reconstructs corrupted masks to restore their original shape and uses these reconstructions as pGTs for QC; and **(3) Wang et al.** [27], a Variational Autoencoder that processes the channel-wise concatenation of image-segmentation pairs, and adjusts their compressed embeddings in the appropriate latent-space using a stochastic iterative search.

Results. The performances of all models are reported in Table 1. The results highlight nnQC's ability to predict quality scores that align closely with GT-based evaluations. Baseline models show inconsistent performance across datasets, struggling with failed segmentations due to their lack of robustness. Figure 3 demonstrates this for Galati et al.[9], which fails to reconstruct degraded prostate segmentations, leading to pseudo-scores resembling random guesses. While this occasionally results in high DSC correlation, it is often an artifact of randomly generated pGT masks. The HD95 correlation further confirms that DSC alone may be insufficient for a complete assessment. Figure 3 also illustrates nnQC's ability to generate anatomically plausible pGTs, achieved through its ToE module, making it a robust solution for segmentation quality evaluation. Across all functions, the high-quality pGTs confirm nnQC's superior performance, particularly in underrepresented instances like the LA cardiac dataset. A major limitation can be foud in the Infant skull-stripping task, whre nnQC

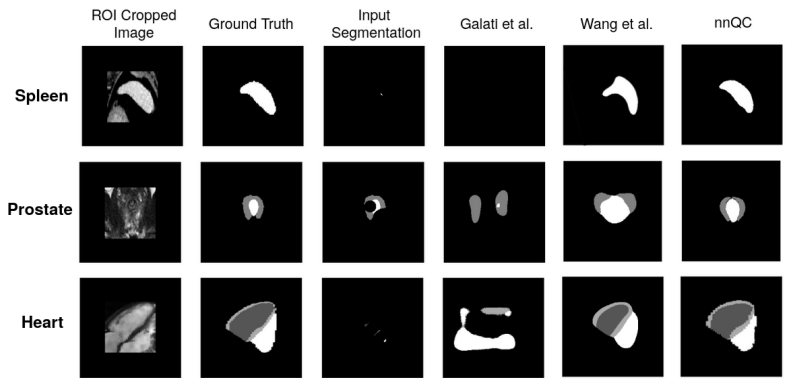

Fig. 3. Examples of pGTs from very poor organ segmentations. nnQC generates plausible pGTs even in highly degraded cases.

maintains a consistent skull perimeter, even simplifying internal details. In contrast, for Fournel et al. [8], we set the HD95 correlation to *N/A*, as it is unbounded and unsuitable for simple regression models. Their method struggles across datasets, achieving an average DSC correlation of 0.42 ± 0.04, though outperforming Galati et al. (0.11 ± 0.30). For Wang et al., we obtained better performance, scoring an average DSC correlation of 0.69 ± 0.24, yet underperforming our method. However, their performance improves on datasets with low mask variability, such as the cardiac dataset M&M2 that comprises the LA view which consists in just one slice per subject. This is reflected in Figure 3.

Implementation Details. We train nnQC on 2D slices, and aggregate the 2D predictions on 3D volumes to compute evaluation metrics. Each volume is scaled according to its modality with the retrieved intensity ranges described in Section 2.4. We take advantage of MONAI and PyTorch for the implementation. Training experiments are run on a 12Gb NVIDIA Titan Xp, and a inference on 80Gb NVIDIA A100.

4 Conclusion

We introduced nnQC, a novel task-agnostic QC framework for segmentation masks, generating reliable pseudo-ground truths (pGTs) through a self-adaptive sampling process. Our *Team of Experts (ToE)* module independently processes image and segmentation data, dynamically balancing their contributions via cross-attention. This enables nnQC to address real-world challenges, including highly degraded segmentations. Furthermore, nnQC extracts dataset-specific *fingerprints* to ensure automatic adaptation across diverse tasks. Evaluations across six organs demonstrated nnQC's superiority over state-of-the-art methods, establishing it as a versatile QC solution. Future work should expand nnQC to additional organs and tumor segmentations and explore its integration as a real-time refinement tool for segmentation correction.

References

1. Audelan, B., Delingette, H.: Unsupervised quality control of image segmentation based on bayesian learning. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22. pp. 21–29. Springer (2019)
2. Bergmann, P., Löwe, S., Fauser, M., Sattlegger, D., Steger, C.: Improving unsupervised defect segmentation by applying structural similarity to autoencoders. arXiv preprint arXiv:1807.02011 (2018)
3. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Transactions on Medical Imaging **37**(11), 2514–2525 (2018)
4. Billot, B., Magdamo, C., Cheng, Y., Arnold, S.E., Das, S., Iglesias, J.E.: Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain mri datasets. Proceedings of the National Academy of Sciences **120**(9), e2216399120 (2023)
5. Campello, V.M., Gkontra, P., Izquierdo, C., Martin-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., et al.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. IEEE Transactions on Medical Imaging **40**(12), 3543–3554 (2021)
6. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 357–366 (2021)
7. Fernandez, V., Pinaya, W.H.L., Borges, P., Graham, M.S., Tudosiu, P.D., Vercauteren, T., Cardoso, M.J.: Generating multi-pathological and multi-modal images and labels for brain mri. Medical Image Analysis **97**, 103278 (2024)
8. Fournel, J., Bartoli, A., Bendahan, D., Guye, M., Bernard, M., Rauseo, E., Khanji, M.Y., Petersen, S.E., Jacquier, A., Ghattas, B.: Medical image segmentation automatic quality control: A multi-dimensional approach. Medical Image Analysis **74**, 102213 (2021)
9. Galati, F., Zuluaga, M.A.: Efficient model monitoring for quality control in cardiac image segmentation (2021)
10. Gur, S., Benaim, S., Wolf, L.: Hierarchical patch vae-gan: Generating diverse videos from a single sample. Advances in Neural Information Processing Systems **33**, 16761–16772 (2020)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
12. Hoopes, A., Mora, J.S., Dalca, A.V., Fischl, B., Hoffmann, M.: Synthstrip: skull-stripping for any brain image. NeuroImage **260**, 119474 (2022)
13. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: Self-adapting framework for deep learning-based bioedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
14. Kalkhof, J., Mukhopadhyay, A.: M3d-nca: Robust 3d segmentation with built-in quality control. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 169–178. Springer (2023)
15. Karani, N., Erdil, E., Chaitanya, K., Konukoglu, E.: Test-time adaptable neural networks for robust medical image segmentation. Medical Image Analysis **68**, 101907 (2021)

16. Kohlberger, T., Singh, V., Alvino, C., Bahlmann, C., Grady, L.: Evaluating segmentation error without ground truth. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 528–536. Springer (2012)
17. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 1–9 (2024)
18. Ma, J., Zhang, Y., Gu, S., An, X., Wang, Z., Ge, C., Wang, C., Zhang, F., Wang, Y., Xu, Y., et al.: Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. Medical Image Analysis **82**, 102616 (2022)
19. Martín-Isla, C., Campello, V.M., Izquierdo, C., Kushibar, K., Sendra-Balcells, C., Gkontra, P., Sojoudi, A., Fulton, M.J., Arega, T.W., Punithakumar, K., et al.: Deep learning segmentation of the right ventricle in cardiac mri: the m&ms challenge. IEEE Journal of Biomedical and Health Informatics **27**(7), 3302–3313 (2023)
20. Rebain, D., Matthews, M.J., Yi, K.M., Sharma, G., Lagun, D., Tagliasacchi, A.: Attention beats concatenation for conditioning neural fields. arXiv preprint arXiv:2209.10684 (2022)
21. Robinson, R., Oktay, O., Bai, W., Valindria, V.V., Sanghvi, M.M., Aung, N., Paiva, J.M., Zemrak, F., Fung, K., Lukaschuk, E., Lee, A.M., Carapella, V., Kim, Y.J., Kainz, B., Piechnik, S.K., Neubauer, S., Petersen, S.E., Page, C., Rueckert, D., Glocker, B.: Real-time prediction of segmentation quality. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. pp. 578–585. Springer International Publishing, Cham (2018)
22. Robinson, R., Valindria, V.V., Bai, W., et al.: Automated quality control in image segmentation: application to the uk biobank cardiovascular magnetic resonance imaging study. Journal of Cardiovascular Magnetic Resonance **21**(1), 18 (2019)
23. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
24. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
25. Saha, A., Twilt, J.J., Bosma, J.S., van Ginneken, B., Yakar, D., Elschot, M., Veltman, J., Fütterer, J., de Rooij, M., Huisman, H.: Artificial Intelligence and Radiologists at Prostate Cancer Detection in MRI: The PI-CAI Challenge (Study Protocol) (2022)
26. Valindria, V.V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B.: Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. IEEE transactions on medical imaging **36**(8), 1597–1606 (2017)
27. Wang, S., Tarroni, G., Qin, C., Mo, Y., Dai, C., Chen, C., Glocker, B., Guo, Y., Rueckert, D., Bai, W.: Deep generative model-based quality control for cardiac mri segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23. pp. 88–97. Springer (2020)
28. Yang, Q., Yan, P., Zhang, Y., Yu, H., Shi, Y., Mou, X., Kalra, M.K., Zhang, Y., Sun, L., Wang, G.: Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. IEEE transactions on medical imaging **37**(6), 1348–1357 (2018)
29. Zhou, S., Greenspan, H., Davatzikos, C., Duncan, J.S., van Ginneken, B., Madabhushi, A., Prince, J.L., Rueckert, D., Summers, R.M.: A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proc IEEE Inst Electr Electron Eng (2021)