applied sciences

MDPI

Article

Dealing with Unreliable Annotations: A Noise-Robust Network for Semantic Segmentation through A Transformer-Improved Encoder and Convolution Decoder

Ziyang Wang * and Irina Voiculescu

Department of Computer Science, University of Oxford, Oxford OX1 3QG, UK; irina@cs.ox.ac.uk
* Correspondence: ziyang.wang@cs.ox.ac.uk

Abstract: Conventional deep learning methods have shown promising results in the medical domain when trained on accurate ground truth data. Pragmatically, due to constraints like lack of time or annotator inexperience, the ground truth data obtained from clinical environments may not always be impeccably accurate. In this paper, we investigate whether the presence of noise in ground truth data can be mitigated. We propose an innovative and efficient approach that addresses the challenge posed by noise in segmentation labels. Our method consists of four key components within a deep learning framework. First, we introduce a Vision Transformer-based modified encoder combined with a convolution-based decoder for the segmentation network, capitalizing on the recent success of self-attention mechanisms. Second, we consider a public CT spine segmentation dataset and devise a preprocessing step to generate (and even exaggerate) noisy labels, simulating real-world clinical situations. Third, to counteract the influence of noisy labels, we incorporate an adaptive denoising learning strategy (ADL) into the network training. Finally, we demonstrate through experimental results that the proposed method achieves noise-robust performance, outperforming existing baseline segmentation methods across multiple evaluation metrics.

Keywords: image segmentation; noisy label; computed tomography; Vision Transformer

Citation: Wang, Z.; Voiculescu, I. Dealing with Unreliable Annotations: A Noise-Robust Network for Semantic Segmentation through A Transformer-Improved Encoder and Convolution Decoder. *Appl. Sci.* **2023**, *13*, 7966. https://doi.org/10.3390/app13137966

Academic Editor: Valentino Santucci

Received: 28 May 2023
Revised: 22 June 2023
Accepted: 26 June 2023
Published: 7 July 2023

1. Introduction

Encoder–decoder network architecture has emerged as a dominant design in medical image segmentation, starting with U-Net, a fully symmetric variant [1]. The encoder in this architecture primarily focuses on extracting pixel location features via convolutional operations and downsampling. The decoder, conversely, reinstates spatial dimensions and pixel location information using deconvolutional operations. The copy-and-crop connection that exists between encoder and decoder layers facilitates the seamless transfer of multiscale semantic information. With this advanced framework, U-Net has been successfully deployed for segmenting a variety of regions within the human body, proving its efficacy across CT [2–4], ultrasound [5–8], and MRI modalities [9–11].

The segmentation capabilities of U-Net were later expanded upon by Çiçek, who introduced a 3D version capable of achieving volumetric segmentation through the extraction of sparsely annotated volumetric images [12]. Another significant advancement was made by Oktay, who introduced an attention gate model. This model enabled CNNs to automatically learn the structures of targets with diverse shapes and sizes [11]. The resultant attention U-Net-based model demonstrated superior sensitivity and accuracy while minimizing computational overhead. Kolarik also contributed to the development of the model by proposing a high-resolution Dense-UNet by comparing the performance of 2D and 3D convolutional operations within residual networks and densely connected networks, respectively [13].

Ongoing research related to various U-Net architectures such as residual networks, inception networks, densely connected networks, and feature normalization continues to

advance the field [2,14–16]. The CNN-based U-Net has been ubiquitous in its application, which has led to the creation of a family of CNN-based U-Nets [2,3,13–15,17–19].

Vision Transformers (ViT) have garnered considerable attention as an influential architecture for diverse computer vision tasks, surpassing traditional CNNs in many instances [20]. Distinct from CNNs, which employ local convolutions to capture spatial information, ViTs utilize self-attention mechanisms [21] to simultaneously process both local and global contextual information. This attribute enables ViT networks to effectively capture intricate patterns and details, rendering them particularly suitable for medical image segmentation tasks that require a sufficient accurate delineation of complex structures [9,22–25]. Current ViT-improved segmentation networks in the literature are largely designed around a U-Net style architecture with self-attention layers. TransUNet [9] was introduced to address the bottleneck of UNet with ViT layers when processing high-level feature information. SwinUNet [22] was designed to utilize a promising shift-window-modified ViT [26] to construct a pure ViT-based UNet-style segmentation network. UNETR [23] introduced explored the ViT-based layer to tackle 3D medical semantic segmentation tasks.

The remarkable performance of both ViT- and CNN-based networks is possible mainly due to the availability of large volumes of high-quality data annotations. This need also constitutes a barrier to deploying advanced machine learning networks to clinical contexts.

Semisupervised learning (SSL) [6,24,27,28], and weakly supervised learning (WSL) [29–31], which aim to tackle the expensive cost of labeling data, have been explored to ViT and CNN, but their application in medical image segmentation still remains relatively underexplored, particularly in the context of noisy labels. The term noisy labels refers to inaccuracies and inconsistencies in the annotated ground truth data, and can significantly challenge network training and generalization [17,32,33]. In medical image segmentation, noisy labels are often inevitable due to factors such as interobserver variability, lack of experience of junior clinicians, and the sheer intricate nature of anatomical structures.

Figure 1 illustrates example CT spine images, corresponding ground truth segmentation masks, and simulated noisy labels. The annotation process may not output masks at a gold-standard level, but may rather generate labeled features that exhibit erosion or dilation of ideal contours, along with various elastic transformations. We refer to these alterations as noisy labels and conjecture their potential detrimental influence on the model's performance. In actual clinical settings, these deviations from the gold standard are likely to be less dramatic.

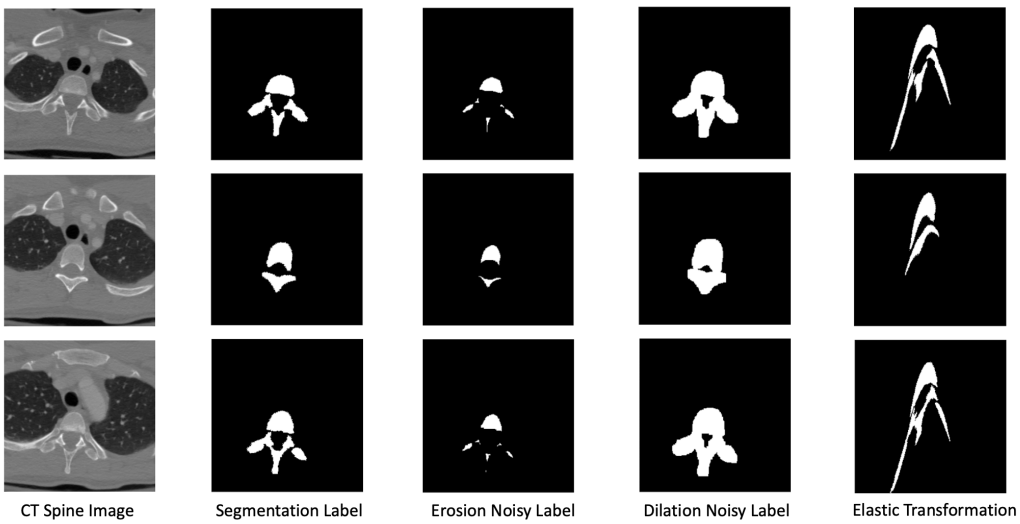

Figure 1. The example images of CT spine images and corresponding segmentation ground truth and noisy labels generated by erosion, dilation, and elastic transform.

By incorporating denoising techniques into the segmentation network, it becomes possible to develop noise-robust networks capable of effectively segmenting medical images in the presence of unreliable annotations. Consequently, this could lead to improved segmentation accuracy and enhanced clinical applicability, contributing to advancements in medical image analysis and diagnosis. Further investigation of ViT-based architectures, combined with sophisticated denoising techniques, may yield transformative outcomes in medical image segmentation, overcoming the limitations posed by noisy labels and facilitating more accurate and reliable diagnostic tools. The detailed noise-robust ViT-modified encoder–CNN decoder network proposed in this paper, named NR-UNet, is introduced in Section 2. The demonstration code is accessible at https://github.com/ziyangwang007/VIT4UNet.

There are four categories of contributions to this work:

1. Inspired by the previous success of CNN and ViT, a ViT-based modified encoder and CNN-based decoder UNet-style segmentation network is proposed.
2. To simulate real clinical scenarios, noise is manually added to the ground truth to create noisy labels, constructing a CT spine segmentation dataset with noisy labels for evaluation purposes.
3. A simple and efficient adaptive denoising learning (ADL) is proposed for segmentation network to achieve a noise-robust segmentation framework.
4. The proposed segmentation network with ADL attains competitive performance against other baseline methods across various evaluation metrics under the same data conditions with both accurate and noisy label sets.

2. Methods

The architecture of NR-UNet is depicted in Figure 2. This symmetrical architecture comprises encoders and decoders, denoted as En_i and De_i, respectively, where $i \in [1, 2, 3, 4]$ indicates the level of encoders and decoders. The transfer of feature maps is conducted between each level of *En* and *De*, following the approach proposed by [1]. The CNN-based en/decoders consist of two-layer CNNs and batch normalization (BN) with up- or downsampling. The ViT-based encoders are designed to extract hidden high-level feature information in En_4 and use a bottleneck structure to further enhance feature learning, and 2-layer ViT is designed in each block. The ViT layer includes layer normalization (LN) [34], multihead self-attention (MSA), and multilayer perceptron (MLP), as described in [21]. The proposed ADL is briefly sketched in Figure 3, where a certain number of noisy labels are able to be detected and removed. We provide details on the segmentation network and adaptive denoising learning strategy in the sections below.

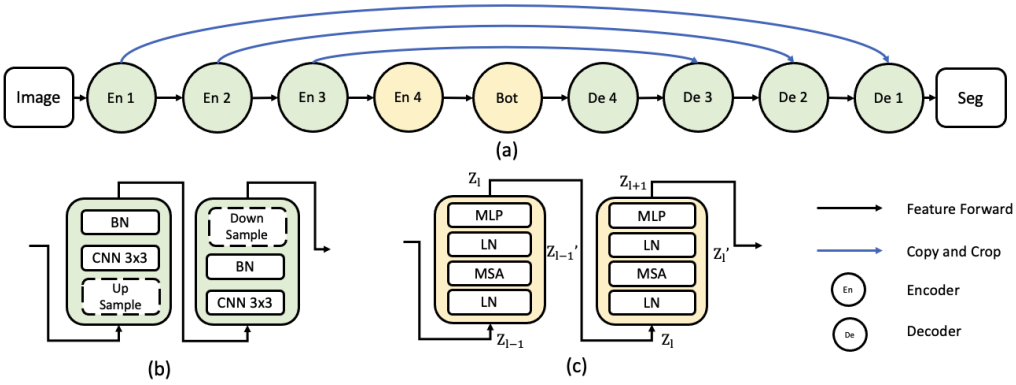

Figure 2. The framework of the proposed NR-UNet. (**a**) The U-shaped encoder–decoder segmentation network through transformer-improved encoder and convolution decoder. ViT-based blocks and CNN-based blocks are green and yellow, respectively. (**b**) The green CNN-based block consists of two successive CNN layers. (**c**) The yellow ViT-based block consists of two successive ViT layers.

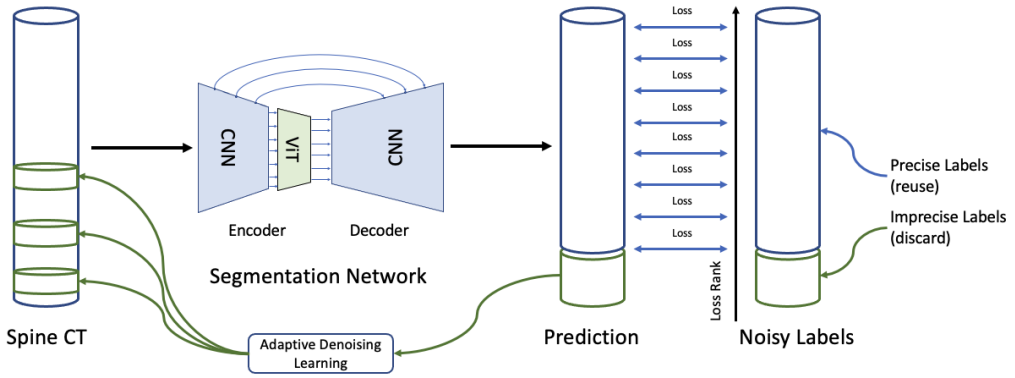

Figure 3. The adaptive denoising learning strategy during training. Predictions get generated through conventional training. Labels with higher prediction losses are more likely to be considered as noisy. ADL detects and removes a specific number of labels deemed as noisy during training. Precise (in blue) and imprecise (in green) labels get classified and, if imprecise, discarded.

2.1. Vision Transformer Layer

In a ViT layer, an input feature map is first partitioned into a fixed number of non-overlapping patches. These patches are subsequently linearly embedded into a flattened vector, which serves as the input token for the transformer. A learnable positional encoding is added to the input token to incorporate spatial information. The encoded input tokens are then fed through a series of ViT layers, each comprising a MSA and a MLP. The MSA enables the model to capture long-range dependencies among input tokens, while the MLP refines the features extracted by the MSA. LN is applied within the ViT layers to stabilize the training process and promote model convergence. The basic ViT-based layer pipeline is shown in Figure 2c, and is described by the following six steps:

(i) The process of tokenization converts the input image x of dimensions $H \times W$ into a sequence of flattened 2D patches denoted $\{x_i^p \in \mathbb{R}^{P^2 \cdot C}\}_{i=1,\ldots,N}$. Each patch measures $P \times P$, and $N = \frac{H \times W}{P^2}$ accounts for the total number of image patches, thereby defining the input sequence length.

(ii) These patches are subsequently mapped onto vectors x^p in a latent D-dimensional embedding space using a trainable linear projection. To capture the spatial information inherent in the patches, learnable position embeddings are added to the patch embeddings, as follows:

$$z^0 = [x_1^p E; x_2^p E; \ldots; x_N^p E] + E_{\text{pos}} \tag{1}$$

In this expression, $E \in \mathbb{R}^{(P^2 \cdot C) \times D}$ represents the patch embedding projection, $E_{\text{pos}} \in \mathbb{R}^{N \times D}$ is the position embedding, and z^0 provides the feature map input to the first ViT layer.

(iii) Comprising L layers, each incorporating an MSA and a MLP, the transformer encoder's output from the l^{th} layer is illustrated as

$$z_l^{'} = \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1} \tag{2}$$

$$z_l = \text{MLP}(\text{LN}(z_l^{'})) + z_l^{'} \tag{3}$$

where $\text{LN}(\cdot)$ stands for the layer normalization operator, and z_l stands for the encoded image.

(iv) The MLP is a fully connected feedforward neural network that consists of multiple layers of nodes. In the proposed NR-UNet, the MLP refines the features extracted

by the MSA mechanism. The MLP involves two linear layers interspersed with a GELU activation function [35], defined as

$$\mathrm{MLP}(z_l^{'}) = \mathrm{Linear}_2(\mathrm{GELU}(\mathrm{Linear}_1(z_l^{'}))) \tag{4}$$

In this equation, Linear_1 and Linear_2 denote the first and second linear layers. The subsequent MSA is composed of multiple self-attention heads that operate in parallel to capture different aspects of the input tokens. Each self-attention head calculates attention scores employing Query (Q), Key (K), and Value (V) matrices, which are derived from the input tokens via linear transformations:
The matrices Query (Q), Key (K), and Value (V) are derived from the input tokens $z_l^{'}$ through linear transformations:

$$Q = z_l^{'} W_Q, \quad K = z_l^{'} W_K, \quad V = z_l^{'} W_V \tag{5}$$

where W_Q, W_K, and W_V are learnable weight matrices [21].

(v) The attention scores are computed by taking the dot product of the Q and K matrices, subsequently scaling and normalizing through softmax:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{6}$$

where d_k denotes the dimensionality of the Key vectors.

(vi) The individual outputs of the self-attention heads are concatenated and linearly transformed to generate the final output of the MSA:

$$\mathrm{MSA}(z_l^{'}) = \mathrm{Concat}(\mathrm{Head}_1, \ldots, \mathrm{Head}_H)W_O \tag{7}$$

where Head_i is the output of the i-th self-attention head, H is the total number of heads, and W_O is another learnable weight matrix.

2.2. ViT Encoder to CNN Decoder

Drawing inspiration from [1,9,22], our proposed NR-UNet architecture combines a ViT-improved encoder with a CNN-based decoder into a hybrid segmentation network. The model initially employs a CNN for feature extraction, thereby creating an insightful feature map from the input image. However, instead of using raw images directly, it performs patch embedding on 1×1 patches extracted from the CNN feature map. This methodological variation amplifies the importance of intermediate high-resolution CNN feature maps within the decoder pathway.

As outlined in [9], our model utilizes a cascaded upsampling path (CUP), which involves a series of upsampling actions to decode the concealed features, ultimately generating the final segmentation mask. The hidden feature sequence $z_L \in \mathbb{R}^{HW/P^2 \times D}$ is first reshaped to $\frac{H}{P} \times \frac{W}{P} \times D$ dimensions, after which we construct the CUP by progressively stringing together several upsampling blocks. The purpose of these blocks is to methodically achieve full resolution, moving from $\frac{H}{P} \times \frac{W}{P}$ to $H \times W$. Each upsampling block consists of a $2\times$ upsampling operator, a 3×3 convolution layer, and a ReLU layer, executed in that order.

2.3. Noisy Labels

In order to evaluate the effectiveness of our proposed ADL in handling noisy labels, we simulate a scenario with imprecise annotations by introducing artificial noise into a dataset whose initial annotations are considered correct. The process of simulating noisy labels in our dataset is as follows:

a Starting with a dataset containing perfect annotations, we randomly select a subset of annotations to be altered, and the ratio of noisy labels to the whole dataset is denoted as $\beta \in [0, 1]$.

b For the selected annotations, we apply image-processing operations such as erosion, dilation, and elastic transformation to generate noisy labels. These operations mimic the types of noise that could be present in real-world clinical scenarios, where annotations might be imperfect due to various factors. Erosion and dilation are fundamental morphological operations. Let A be the binary annotation mask and B be a structuring element. The erosion ($\ominus$) and dilation ($\oplus$) operations can be defined as

$$(A \ominus B)(x, y) = \min_{(i,j) \in B} A(x - i, y - j) \tag{8}$$

$$(A \oplus B)(x, y) = \max_{(i,j) \in B} A(x + i, y + j) \tag{9}$$

where (x, y) represents the pixel coordinates in the image.
Elastic transformation is a nonlinear deformation technique that can simulate the local warping of shapes. Given an image $I(x, y)$ and two displacement fields $\Delta x(x, y)$ and $\Delta y(x, y)$, which are generated by Gaussian smoothing of random fields, an elastic transformation can be defined as

$$I_{\text{transformed}}(x, y) = I(x + \alpha \Delta x(x, y), y + \alpha \Delta y(x, y)) \tag{10}$$

where α is the deformation scale factor.

c The altered annotations are then used to replace their corresponding original annotations in the dataset, creating a new dataset with a mix of perfect and noisy labels. By simulating noisy labels in this manner, we create a dataset that can challenge experiments, enabling us to assess the noise-robustness of our proposed method and compare its performance against existing high-performing techniques.

2.4. Adaptive Denoising Learning Strategy

In real-world applications, labels are inherently noisy, in the sense that they do not represent a precise ground truth. Our adaptive denoising learning (ADL) strategy is fairly straightforward, and yet efficient. To simulate different scenarios involving noise labels, a proportion β of the provided ground truth masks in the training data (which are accurate) are replaced with noisy labels generated synthetically, as described in Section 2.3. These labels exhibit erosions, dilations, or elastic transformations, which are illustrated in Figure 1.

Drawing inspiration from O2U-Net [36], in every training epoch, we log the discrepancy between each prediction and its corresponding ground truth label. Labels accruing higher losses are more likely to be considered noisy, and are subsequently discarded at the epoch's conclusion.

During the training process, our ADL strategy identifies and discards a specific quantity of labels with high loss values. As illustrated in Figure 3, a considerable quantity of noisy labels are detected and eliminated at the beginning of the training iteration; this quantity decreases gradually, as the training progresses. This reflects the transition of the training process from a state of underfitting to overfitting [17]. The number $N(t)$ of labels identified and removed at each epoch is calculated as

$$N(t) = \frac{\beta y}{k^2}(k - t) \tag{11}$$

In this equation, t corresponds to the current training epoch, β represents the proportion of items in the training dataset subjected to noise, k stands for the total count of training epochs, and y indicates the total count of masks. By integrating ADL into the training process, we amplify the network's robustness to noisy labels. This strategy, in turn, improves the model's performance and clinical applicability.

3. Experiments and Results

3.1. Dataset

Our research makes use of a publicly available spine dataset, courtesy of the University of California Irvine Medical Centre's School of Medicine and the National Institutes of Health [37]. This dataset includes CT scans from 10 individuals aged from 16 to 35 years. Each scan incorporates up to 600 slices of resolution 512 × 512 and interslice spacing of 1 mm. All images are normalized and resized to 256 × 256 pixels for uniformity. The dataset provides ground truth masks for each image, and a portion of these masks is subjected to noise (Section 2.3). Out of the 10 scans, 9 were used for training, which produced a total of 5043 images. The last scan, consisting of 552 images, was set aside for testing. Validation was conducted on 10% of the training data, ensuring no overlap occurred between the training and testing datasets.

3.2. Experimental Setup

The proposed NR-UNet was implemented using Python and Tensorflow, with experiments carried out on a system equipped with an Nvidia GeForce RTX 3090 GPU, with 24 GB memory, and an Intel CPU i9-10900K. In this experimental configuration, we utilized an input feature map of size $(256, 256, 1)$, with the number of CNNs on each layer of decoders set to 64, 128, 256, 512 for all U-shape encoder–decoder networks, incorporating two successive CNN layers within each encoder/decoder. With regards to the ViT-related design elements, we assigned an image patch-embedding dimension of 768, an MLP count of 1024, and 12 heads for the MSA. The bottleneck of the NR-UNet contained six ViT layers. Given that the downstream task was 2D image semantic segmentation (i.e., a per-pixel binary classification), the final layer was a 1 × 1 CNN layer. Runtimes for 50 epochs, including data transfer, ranged between 500 and 800 min. We opted for a training batch size of 4, and utilized the Adam optimizer, with a learning rate of 10^{-5}. The chosen loss function relied on the Dice coefficient (Equation (12)), a frequently used metric for assessing overlap in semantic segmentation, particularly suited to managing the imbalance between a region of interest (ROI) and the background.

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \cdot \sum_{i=1}^{N} y_i \hat{y}i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i + \varepsilon} \tag{12}$$

where ε is a small positive constant (e.g., 10^{-8}) added to the denominator to prevent division by zero, y_i denotes the ground truth mask, $\hat{y}_i$ represents the predicted segmentation mask, and N is the total number of pixels in the image. This experimental setup allowed us to thoroughly investigate the contribution of the proposed adaptive denoising learning strategy to the overall performance of the NR-UNet in the context of medical image segmentation.

3.3. Metrics

Beyond the Dice coefficient, we evaluated the performance of NR-UNet using an assortment of standard overlap metrics: intersection over union (IoU), accuracy (Acc), precision (Pre), recall (Rec), and specificity (Spe). These evaluation measures are detailed in Equation (13), where TP, TN, FP, and FN represent the number of pixels classified as true positive, true negative, false positive, or false negative. The IoU measures the extent of the overlap between the predictions of our model and the ground truth; accuracy evaluates the overall correctness of the model's predictions, with high accuracy in a medical context potentially signifying a model's efficacy in distinguishing between different classes—a crucial factor for disease detection and treatment planning; precision and recall are crucial performance indicators that reflect a model's proficiency in accurately identifying positive cases and evading false negatives, respectively. In the realm of medical image analysis, high precision ensures that positive detections are indeed relevant, thereby mitigating risks associated with false positives. Simultaneously, high recall improves the model's capability to spot all areas of concern, minimizing overlooked detections. Specificity, on the

other hand, quantifies the model's capacity to correctly identify negative cases. Clinically, high specificity can curtail false-positive rates, thus preventing patients from undergoing unwarranted further examinations or treatments. Through the implementation of these diverse metrics, our intention is to offer an exhaustive evaluation of the proposed model's performance, substantiating its potential value in clinical scenarios.

$$\begin{aligned} Dice &= \frac{2TP}{2TP + FP + FN} \\ IoU &= \frac{TP}{TP + FP + FN} \\ Accuracy &= \frac{TP + TN}{TP + TN + FP + FN} \\ Precision &= \frac{TP}{TP + FP} \\ Recall &= \frac{TP}{TP + FN} \\ Specificity &= \frac{TN}{TN + FP} \end{aligned} \tag{13}$$

3.4. Experiments in a Noise-Free Setup

Figure 4 illustrates eight examples of raw images, and the predicted result of a number of different networks, all training on accurate labels: UNet [1], residual UNet [13], dense UNet [13], MultiResUnet [19], LinkNet [38], FPN [39], UNet++ [2], UNet3+ [18], VNet [10], RARUNet [17], and QAPNet [3]. Our NR-UNet was compared with these for a setting of $\beta = 0$ (hence, also training on accurate labels).

Batch normalization [40] and Dropout [41] were deployed after CNN layers for all UNet-based baseline methods for a fair comparison. A quantitative comparison of NR-UNet against the other ten baseline methods is given in Table 1. The best results are highlighted in bold.

Table 1. Direct comparison against existing algorithms.

Network ($\beta = 0$)	Dice	IoU	Acc	Pre	Rec	Spe	Par$_{10^6}$
FPN [39]	0.9373	0.8821	0.9944	0.9191	0.9563	0.9961	17.59
Residual UNet [13]	0.9416	0.8897	0.9949	0.9481	0.9353	0.9976	9.90
VNet [10]	0.9446	0.8950	0.9950	0.9202	**0.9703**	0.9961	14.74
LinkNet [38]	0.9524	0.9091	0.9959	0.9662	0.9390	0.9985	20.32
UNet [1]	0.9580	0.9193	0.9963	0.9619	0.9541	0.9983	8.64
Dense UNet [13]	0.9612	0.9252	0.9966	0.9600	0.9624	0.9982	15.47
MultiRes UNet [19]	0.9644	0.9312	0.9969	0.9633	0.9655	0.9983	7.76
UNet++ [2]	0.9659	0.9340	0.9970	0.9676	0.9642	0.9985	8.86
RARUNet [17]	0.9674	0.9369	0.9972	0.9721	0.9629	0.9987	11.79
QAPNet [3]	0.9690	0.9399	0.9973	0.9715	0.9666	0.9987	15.14
NR-UNet	**0.9703**	**0.9424**	**0.9974**	**0.9740**	**0.9667**	**0.9988**	182.90

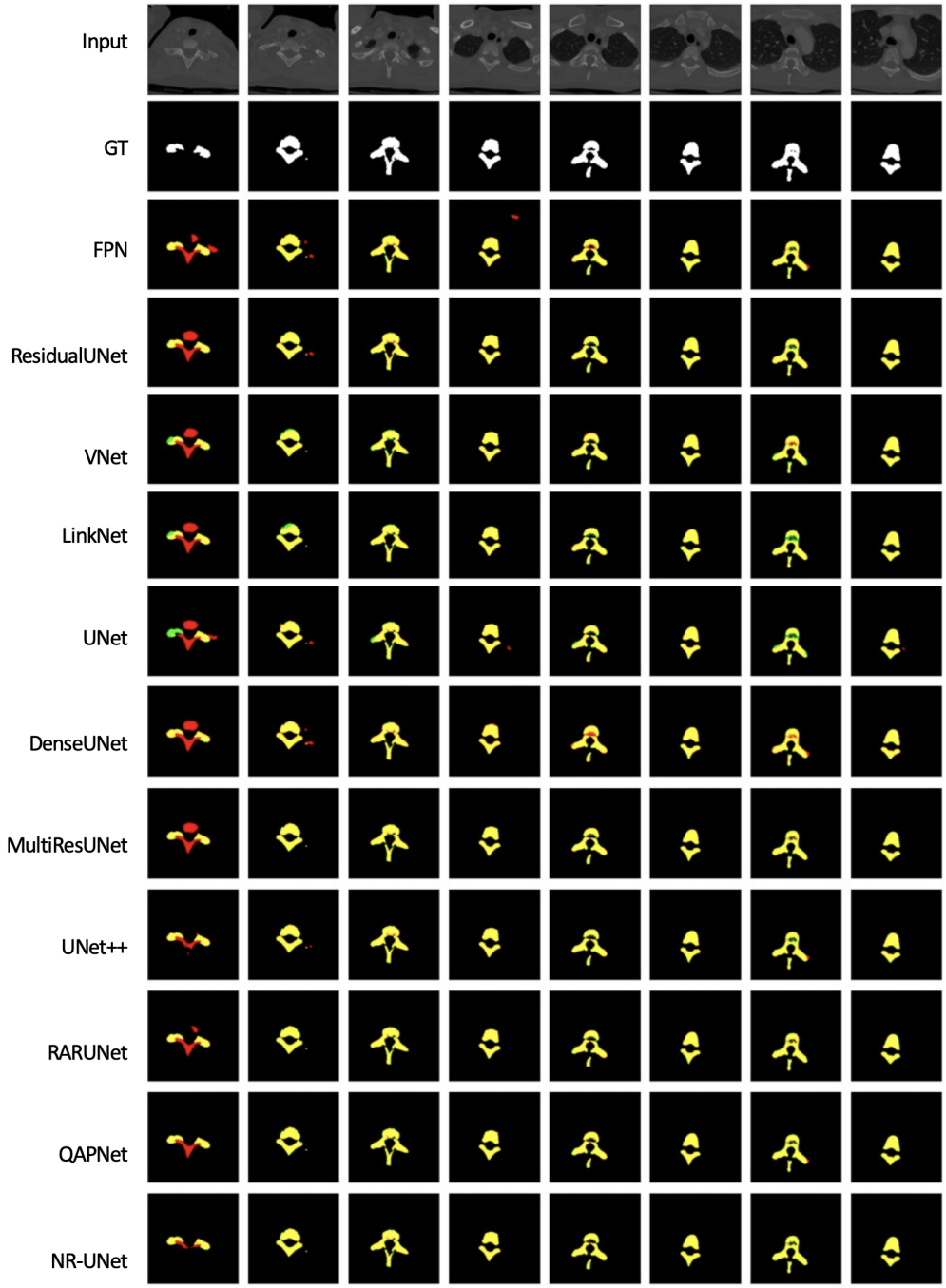

Figure 4. The example images of input CT spine images; prediction of each network against ground truth.

In Figure 5, we present the box plot of the Dice coefficient for each model prediction. The Dice coefficient quantifies the performance of each model by assessing the overlap between the model prediction and the ground truth. The coefficients range between 0 and 1, where 1 signifies a perfect overlap between the prediction and the ground truth, and 0 signifies no overlap. The variation in box heights and positions provides an overview of the performance differences across the models. A higher median and a narrower box suggest that a model has a high overall performance and a consistent prediction quality across different slices. Additionally, pairwise *t*-tests were conducted to determine the statistical significance of the differences in Dice coefficients between NR-UNet and UNet, where the *t*-statistic is 2.64 and the *p*-value is 0.0085. A low *p*-value (typically < 0.05) indicates a statistically significant difference in performance between two models. A comparison

of the training progress of NR-UNet against UNet is depicted in Figure 6. The line charts illustrate the evolution of loss on the training set (in blue) and the validation set (in red). On the y-axis, each type of loss is plotted against the number of training epochs on the x-axis.

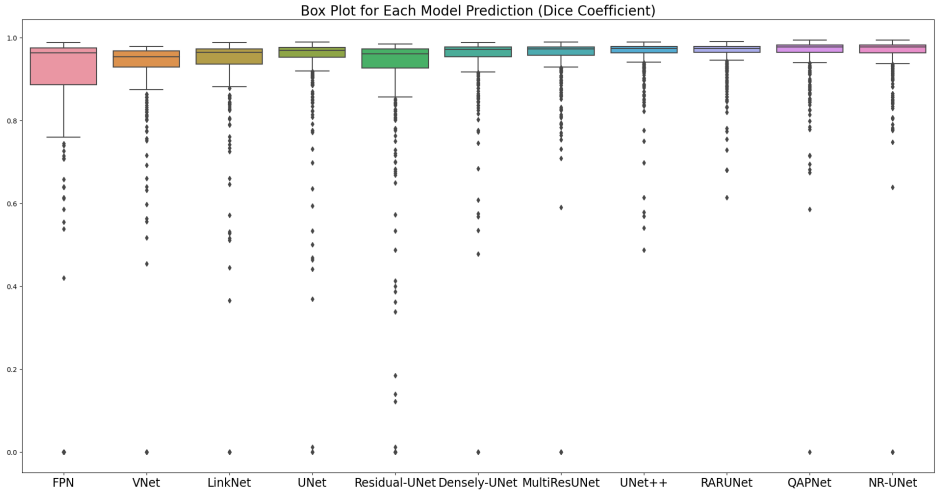

Figure 5. Box plot for the Dice coefficient distribution of prediction by each model.

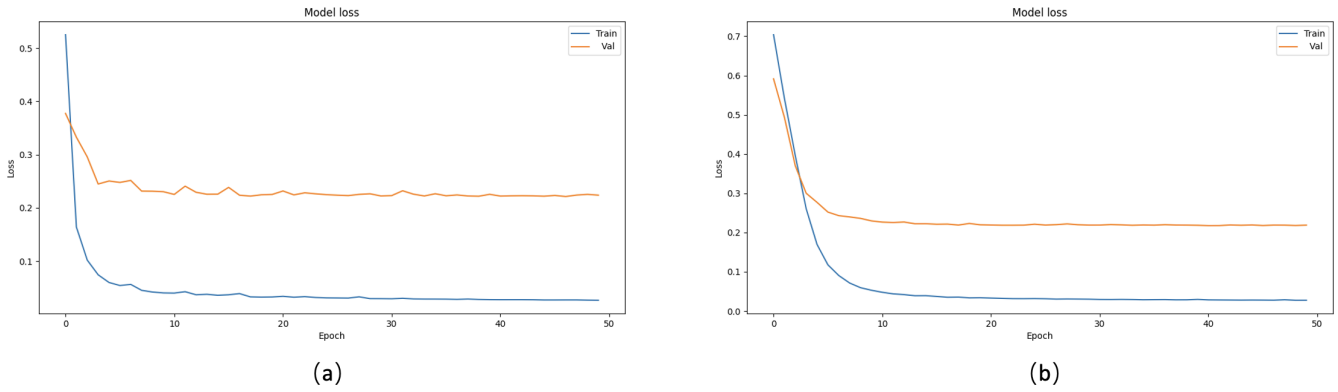

Figure 6. History of training loss (in blue) and validation loss (in red) against the number of epochs for (**a**) the CNN encoder–decoder, and for (**b**) the ViT-improved encoder and CNN decoder.

3.5. Experiments in a Noisy Setup

To evaluate the impact of the proposed ADL contribution, our core experiments assessed the algorithms' robustness to noise in the annotations. Table 2 specifically examines the effect of the ADL strategy. 'Proportion' represents the percentage β of noisy labels introduced into the training dataset, with various other proportions also tested beyond those shown in the table (but omitted here to avoid clutter). The 'Network' column demonstrates that different networks were employed for training separately, and the presence of a module implementing the ADL strategy (annotated with check marks) improves performance across the board. Whilst in Table 1, the various networks performed similarly, it is evident from Table 2 that our proposed ADL strategy mitigates the impact of noisy labels, bringing the segmentation performance to levels that can be considered acceptable in many practical applications. Eight example predictions by NR-UNet under different data situations (i.e., different ratio β of noisy labels) are sketched in Figure 7. In this setup, NR-UNet is noticeably more effective than its counterparts. By comparing the outcomes of different algorithms and varying the proportion of noisy labels introduced into the training dataset, we gain valuable insights into the robustness and adaptability of our method in addressing the challenges posed to medical image segmentation tasks by noisy annotations.

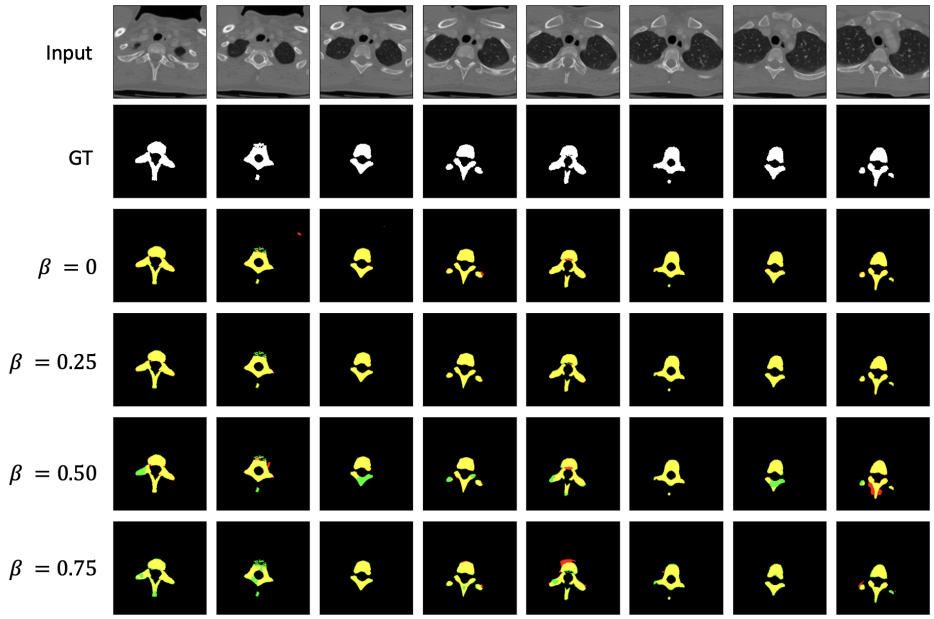

Figure 7. The example images of input CT spine images; prediction of NR-UNet against ground truth under various data situations.

Table 2. Ablation studies of adaptive denoising learning via various segmentation networks under different proportions of noisy labels.

Proportion (β)	Network	ADL	Dice	IoU
75%	Residual UNet	✗	0.7962	0.6614
		✓	0.8210	0.6964
75%	UNet	✗	0.8004	0.6672
		✓	0.8337	0.7148
75%	NR-UNet	✗	0.8196	0.6943
		✓	**0.8466**	**0.7340**
50%	Residual UNet	✗	0.8179	0.6919
		✓	0.8453	0.7321
50%	UNet	✗	0.8188	0.6932
		✓	0.8564	0.7489
50%	NR-UNet	✗	0.8362	0.7185
		✓	**0.8832**	**0.7908**
25%	Residual UNet	✗	0.9002	0.8185
		✓	0.9213	0.8541
25%	UNet	✗	0.9084	0.8322
		✓	0.9303	0.8697
25%	Dense UNet	✗	0.9096	0.8342
		✓	0.9284	0.8664
25%	NR-UNet	✗	0.9101	0.8350
		✓	**0.9532**	**0.9106**

4. Conclusions

Our research into the impact of noise in annotation labels suggests that for certain semantic segmentation tasks, near-perfect hand-drawn contours are not a strict necessity. Through our experiments, we show that the proposed NR-UNet, coupled with the ADL strategy, performs competitively under varying levels of label noise.

While the results from the proposed NR-UNet and ADL strategy are promising, there is room for further improvement. As part of our future work, we plan to broaden the applicability of the NR-UNet and ADL strategy to other medical imaging modalities, such as MRI and ultrasound. We will also assess how generalizable they are across a wider range of clinical applications. These methods can also be used in conjunction with weakly

supervised learning tasks, improving the reliability of safety-critical scenarios like medical screening or diagnosis on the strength of minimal hand-annotation requirements.

Author Contributions: Conceptualization, Z.W. and I.V.; Validation, Z.W.; Investigation, Z.W.; Resources, I.V.; Writing – original draft, Z.W.; Writing—review & editing, Z.W. and I.V.; Supervision, I.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset used in this study is public available at http://spineweb.digitalimaginggroup.ca/

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015 ; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
2. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
3. Wang, Z.; Voiculescu, I. Quadruple augmented pyramid network for multi-class COVID-19 segmentation via CT. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Guadalajara, Mexico, 1–5 November 2021; pp. 2956–2959.
4. Gao, Y.; Guo, J.; Fu, C.; Wang, Y.; Cai, S. VLSM-Net: A Fusion Architecture for CT Image Segmentation. *Appl. Sci.* **2023**, *13*, 4384. [CrossRef]
5. Noble, J.A.; Boukerroui, D. Ultrasound image segmentation: A survey. *IEEE Trans. Med. Imaging* **2006**, *25*, 987–1010. [CrossRef] [PubMed]
6. Wang, Z.; Voiculescu, I. Triple-view feature learning for medical image segmentation. In Proceedings of the Resource-Efficient Medical Image Analysis: First MICCAI Workshop, REMIA 2022. Singapore, 22 September 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 42–54.
7. Benjdira, B.; Ouni, K.; Al Rahhal, M.M.; Albakr, A.; Al-Habib, A.; Mahrous, E. Spinal cord segmentation in ultrasound medical imagery. *Appl. Sci.* **2020**, *10*, 1370. [CrossRef]
8. Wang, Z. Deep learning in medical ultrasound image segmentation: A review. *arXiv* **2020**, arXiv:2002.07703.
9. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
10. Milletari, F.; Navab, N.; Ahmadi, S.A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 fourth international conference on 3D vision (3DV), Stanford, CA, USA , 25–28 October 2016; pp. 565–571.
11. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
12. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Athens, Greece, 17–21 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 424–432.
13. Kolařík, M.; Burget, R.; Uher, V.; Říha, K.; Dutta, M.K. Optimized high resolution 3d dense-u-net network for brain and spine segmentation. *Appl. Sci.* **2019**, *9*, 404. [CrossRef]
14. Guan, S.; Khan, A.A.; Sikdar, S.; Chitnis, P.V. Fully Dense UNet for 2-D Sparse Photoacoustic Tomography Artifact Removal. *IEEE J. Biomed. Health Inform.* **2019**, *24*, 568–576. [CrossRef] [PubMed]
15. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote. Sens.* **2020**, *162*, 94–114. [CrossRef]
16. Li, W.; Tang, Y.M.; Wang, Z.; Yu, K.M.; To, S. Atrous residual interconnected encoder to attention decoder framework for vertebrae segmentation via 3D volumetric CT images. *Eng. Appl. Artif. Intell.* **2022**, *114*, 105102. [CrossRef]
17. Wang, Z.; Zhang, Z.; Voiculescu, I. RAR-U-Net: A residual encoder to attention decoder by residual connections framework for spine segmentation under noisy labels. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 21–25.
18. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.W.; Wu, J. Unet 3+: A full-scale connected unet for medical image segmentation. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 1055–1059.
19. Ibtehaz, N.; Rahman, M.S. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Netw.* **2020**, *121*, 74–87. [CrossRef] [PubMed]

20. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
22. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the Computer Vision–ECCV 2022 Workshops, Tel Aviv, Israel, 23–27 October2022; Part III; Springer: Berlin/Heidelberg, Germany, 2023; pp. 205–218.
23. Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H.R.; Xu, D. Unetr: Transformers for 3d medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 574–584.
24. Wang, Z.; Zheng, J.Q.; Voiculescu, I. An uncertainty-aware transformer for MRI cardiac semantic segmentation via mean teachers. In Proceedings of the Medical Image Understanding and Analysis: 26th Annual Conference, MIUA 2022, Cambridge, UK, 27–29 July 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 494–507.
25. Wang, J.; Zhang, H.; Yi, Z. CCTrans: Improving Medical Image Segmentation with Contoured Convolutional Transformer Network. *Mathematics* **2023**, *11*, 2082. [CrossRef]
26. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 10012–10022.
27. Bortsova, G.; Dubost, F.; Hogeweg, L.; Katramados, I.; De Bruijne, M. Semi-supervised medical image segmentation via learning consistency under transformations. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, 13–17 October 2019; Proceedings, Part VI 22; Springer: Berlin/Heidelberg, Germany, 2019; pp. 810–818.
28. Li, S.; Zhang, C.; He, X. Shape-aware semi-supervised 3D semantic segmentation for medical images. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, 4–8 October 2020; Proceedings, Part I 23; Springer: Berlin/Heidelberg, Germany, 2020; pp. 552–561.
29. Liu, X.; Yuan, Q.; Gao, Y.; He, K.; Wang, S.; Tang, X.; Tang, J.; Shen, D. Weakly supervised segmentation of COVID19 infection with scribble annotation on CT images. *Pattern Recognit.* **2022**, *122*, 108341. [CrossRef] [PubMed]
30. Lee, H.; Jeong, W.K. Scribble2label: Scribble-supervised cell segmentation via self-generating pseudo-labels with consistency. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, 4–8 October 2020; Proceedings, Part I 23; Springer: Berlin/Heidelberg, Germany, 2020; pp. 14–23.
31. Kervadec, H.; Dolz, J.; Wang, S.; Granger, E.; Ayed, I.B. Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision. In Proceedings of the Medical Imaging with Deep Learning, PMLR, Montreal, QC, Canada, 6–8 July 2020; pp. 365–381.
32. Lu, Z.; Fu, Z.; Xiang, T.; Han, P.; Wang, L.; Gao, X. Learning from weak and noisy labels for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 486–500. [CrossRef] [PubMed]
33. Wang, G.; Liu, X.; Li, C.; Xu, Z.; Ruan, J.; Zhu, H.; Meng, T.; Li, K.; Huang, N.; Zhang, S. A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images. *IEEE Trans. Med. Imaging* **2020**, *39*, 2653–2663. [CrossRef] [PubMed]
34. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
35. Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv* **2016**, arXiv:1606.08415.
36. Huang, J.; Qu, L.; Jia, R.; Zhao, B. O2u-net: A simple noisy label detection approach for deep neural networks. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3326–3334.
37. Yao, J.; Burns, J.E.; Munoz, H.; Summers, R.M. Detection of vertebral body fractures based on cortical shell unwrapping. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Nice, France, 1–5 October 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 509–516.
38. Chaurasia, A.; Culurciello, E. Linknet: Exploiting encoder representations for efficient semantic segmentation. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017; pp. 1–4.
39. Kim, S.W.; Kook, H.K.; Sun, J.Y.; Kang, M.C.; Ko, S.J. Parallel feature pyramid network for object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 234–250.
40. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Pmlr, Lille, France, 6–11 July 2015; pp. 448–456.
41. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.