HPS-Net: Multi-Task Network for Medical Image Segmentation with Predictable Performance

Xin Wei 1, Huan Wan 2, Fanghua Ye 1 and Weidong Min 1,*

School of Software, Nanchang University, 235 East Nanjing Road, Nanchang 330047, China; xinwei@ncu.edu.cn (X.W.); fanghuaye@email.ncu.edu.cn (F.Y.)

School of Computer Information Engineering, Jiangxi Normal University, 99 Ziyang Avenue, Nanchang 330022, China; wan-h@ulster.ac.uk

* Correspondence: minweidong@ncu.edu.cn

Abstract: In recent years, medical image segmentation (MIS) has made a huge breakthrough due to the success of deep learning. However, the existing MIS algorithms still suffer from two types of uncertainties: (1) the uncertainty of the plausible segmentation hypotheses and (2) the uncertainty of segmentation performance. These two types of uncertainties affect the effectiveness of the MIS algorithm and then affect the reliability of medical diagnosis. Many studies have been done on the former but ignore the latter. Therefore, we proposed the hierarchical predictable segmentation network (HPS-Net), which consists of a new network structure, a new loss function, and a cooperative training mode. According to our knowledge, HPS-Net is the first network in the MIS area that can generate both the diverse segmentation hypotheses to avoid the uncertainty of the plausible segmentation hypotheses and the measure predictions about these hypotheses to avoid the uncertainty of segmentation performance. Extensive experiments were conducted on the LIDC-IDRI dataset and the ISIC2018 dataset. The results show that HPS-Net has the highest Dice score compared with the benchmark methods, which means it has the best segmentation performance. The results also confirmed that the proposed HPS-Net can effectively predict TNR and TPR.

Keywords: symmetrical structure; medical image; image segmentation; deep learning; CNNs; loss function

1. Introduction

With the development of medical imaging modalities such as ultrasound, X-ray, computed tomography (CT), and magnetic resonance imaging (MRI), much attention is attracted to developing new medical image-processing (MIP) techniques [1]. Among different types of MIP techniques, medical image segmentation has been a hotspot in recent years, which is considered the most crucial step in clinical diagnosis. Moreover, medical image segmentation is also the prerequisite of many downstream clinical tasks. Nevertheless, medical image segmentation is often characterized by uncertainties. These uncertainties include the uncertainty of the plausible segmentation hypotheses and the uncertainty of segmentation performance. These two types of uncertainties affect the effectiveness of the MIS algorithm and affect the reliability of the subsequent medical diagnosis.

Different medical experts may have their style of segmenting lesions from medical images, and from a medical perspective, all their segmentations could be correct [2]. Additionally, in organ abnormality segmentation, a lesion area is clear and distinct, but even the top experts may not be able to judge whether the lesion is cancer tissue or not. These result in the uncertainty of the plausible segmentation hypotheses, which is common in medical imaging. On the other hand, there is uncertainty in the segmentation performance of a segmentation algorithm. A segmentation algorithm may perform well on some medical images while performing poorly on the other images. Failure to detect these poor segmentations may affect the performance of downstream tasks and even lead to misdiagnosis.
in practice. Most automatic segmentation algorithms treat the segmentation target as a one-to-one mapping from the image to the output mask. However, only providing the most likely segmentation hypothesis is not enough for radiologists. Radiologists need all plausible segmentation hypotheses to avoid the uncertainty of the plausible segmentation hypotheses. They also need predictable statistical measures about these hypotheses to assist them in evaluating the segmentation performance. If possible, radiologists even want to use their specified measurement values to generate segmentation hypotheses.

Many studies have been done to solve the uncertainty of the plausible segmentation hypotheses by producing more than one plausible hypothesis from the models. Balaji et al. [3] propose to learn an ensemble of deep models to generate multiple plausible hypotheses. However, the output segmentations generated by the ensembles lack diversity, and the ensembles usually do not learn rare segmentation hypotheses, because their models are trained independently, and the impact of the rare segmentation hypotheses is weakened by the majority. Abner et al. [4] present a max-margin formulation and the oracle set loss to directly model the M-Best prediction problem. Enlightened by Abner et al. [4], Rupprecht et al. [5] and Ilg et al. [6] combined the oracle set loss with a common deep network with M heads to generate M hypotheses. Nevertheless, all the above-mentioned methods can only generate a fixed number of hypotheses and are not graceful to extend to a large number of hypotheses. This problem can be solved by a conditional variational autoencoder (CVAE) [7], as a CVAE can effectively perform probabilistic inference by learning Gaussian latent variables to model complex distributions. With the learned conditional distribution, a CVAE can generate unlimited and diverse segmentation hypotheses. Due to the excellent segmentation performance of the U-Net [8], Simon et al. [9] combined the CVAE with the U-Net. The combined method is called probabilistic U-net, which can produce an unlimited number of segmentation hypotheses while delivering better performance than network ensembles [3] and M-heads [5,6]. Recently, probabilistic hierarchical segmentation (PHiSeg) [10] further extends the prior net and the posterior net of the probabilistic from a single resolution structure to a hierarchical multi-resolution structure and achieves state-of-the-art performance across multiple datasets.

As another type of uncertainty, the uncertainty of segmentation performance has not drawn enough attention. To reduce this type of uncertainty, a feasible method is to provide the corresponding measure predictions along with each segmentation hypothesis. The measures can be precision, accuracy, the true positive rate, the true negative rate, or other measures. These predictive measures can assist radiologists to evaluate the segmentation performance and decide whether the segmentation hypothesis should be accepted. To the best of our knowledge, there exists no other work that has considered providing the corresponding measure predictions along with each segmentation hypothesis; and there exists no other work that has considered generating segmentation hypotheses based on specified measurement values. To fill these two gaps, we propose a hierarchical predictable segmentation network (HPS-Net). An illustration of the application case of the proposed HPS-Net is shown in Figure 1. HPS-Net can learn a complex probability distribution of the samples in the latent space and has the ability to predict the measurement values; therefore, it can generate an unlimited number of segmentation hypotheses along with their measure predictions. From another perspective, HPS-Net can provide the segmentation hypothesis based on a specified measurement value by continually generating segmentation hypotheses until the predicted measurement value is extremely close to the specified value. HPS-Net is a hierarchical multi-task network for medical image segmentation and segmentation performance prediction, which consists of a new network structure, a new loss function, and a cooperative training mode. As a multi-task network, HPS-Net includes four sub-networks, namely, the posterior network, the prior network, the likelihood network, and the measure network. These sub-networks have multiple symmetrical structures in them, which enable the networks to extract the features of images at different scales. To demonstrate the efficacy of HPS-Net, experiments were conducted on two public datasets—the LIDC-IDRI dataset [11] and the ISIC2018 dataset [12]. The
results show that the proposed HPS-Net can provide effective measure predictions while keeping the maximum segmentation performance in terms of segmentation accuracies of multiple annotations and a single annotation per image.

![Figure 1](image.png)

**Figure 1.** An illustration of the application case of the proposed HPS-Net. HPS-Net can generate an unlimited number of segmentation hypotheses along with their measure predictions and can provide the segmentation hypothesis based on a specified measurement value.

The remainder of the article is organized as follows. In Section 2, we review related work including U-Net, probabilistic U-Net, and PHiSeg. We describe problem formulation and the HPS-Net structure in Section 3. In Section 4, we evaluate the performance of HPS-Net on the LIDC-IDRI dataset [11] and the ISIC2018 dataset [12]. Finally, we conclude our article in Section 5.

2. Related Work

Our proposed HPS-NET is derived from PHiSeg, while PHiSeg is derived from the probabilistic U-Net, where U-Net is a milestone in the field of medical image segmentation and is the derivative source of HPS-NET, PHiSeg, and the probabilistic U-Net. To have a better understanding of the proposed method, we review U-Net, the probabilistic U-Net, and PHiSeg in this section.

2.1. U-Net

Due to the outstanding performance of U-Net, many excellent networks [13–15] are based on or sourced from U-Net. U-Net [8] is derived from the fully convolutional network (FCN) [16] and is a convolutional neural network developed for biomedical image segmentation task. Compared with FCN, the structure of U-Net was modified to adapt to fewer training images and to achieve higher segmenting accuracy. The main idea of U-Net is to supplement a usual contracting network by successive layers, where pooling operations are replaced by upsampling operators. Hence, these layers increase the resolution of the output. Moreover, successive convolutional layers can then learn to assemble a precise output based on the information.

Figure 2 illustrates the structure of U-Net. From Figure 2, we can see that U-Net has two symmetric paths, namely, the left contracting path and the right expanding path. The contracting path is made up of repeated convolutional layers, ReLU layers, and Max-pooling layers, which focuses on capturing context information. The right expanding path contains repeated convolutional layers, ReLU layers, and upsampling layers, which enables precise localization. There are some skip connections between these two paths, which can compensate for the loss of boundary pixels caused by the convolution operations in the contracting path.
Figure 2. U-Net structure. Each dark green box corresponds to a multi-channel feature map. Light green boxes represent copied feature maps. The arrows denote the different operations. Up-conv $3 \times 3$ represents upsampling of the feature map followed by a $3 \times 3$ convolution.

Since the input and output are images of the same size, why construct a symmetric structure that includes downsampling and upsampling? The answer can be explained from the theoretical meaning of symmetry principles. The symmetric structure can be divided into two paths—the downsampling path and the upsampling path. The downsampling path can increase the robustness of some small disturbances of the input image, such as image translation and rotation, thereby reducing the risk of overfitting, reducing the amount of calculation, and increasing the size of the receptive field. The main purpose of the upsampling path is actually to transform the features back to the size of the original image. In addition, since a large amount of information is lost in downsampling, some connections are applied between the layers at the same resolution level to supplement the missing information. The above reasons led to the final adoption of a symmetric structure.

2.2. Probabilistic U-Net

U-Net delivers the best segmentation performance at the time and has been applied to many medical image domains. However, U-Net follows a one-to-one I/O mode, so it cannot solve the uncertainty of the plausible segmentation hypotheses. Based on U-Net and the conditional variational auto encoder (CVAE) [7], Simon et al. [9] proposed a segmentation network called probabilistic U-Net, which can output an unlimited number of segmentation hypotheses for an input image. The network structure of the probabilistic U-Net is shown in Figure 3a. Similar to the general auto encoder, the CVAE is also symmetrical and consists of two symmetric sub-networks called prior net and posterior net, which are corresponding to the encoder and decoder in an auto encoder. The inputs of the prior net and the posterior net are the medical image and the ground truth of the segmentation. Different from the general auto encoder, the CVAE assumes that the data are generated by some random processes, and the CVAE performs probabilistic inference by learning Gaussian latent variables to model complex distributions in the latent space. In CVAE, the parameters in the network are assigned the meaning of $\mu$ and $\sigma$ in a Gaussian distribution. During the training process, the CVAE optimizes a large number of Gaussian latent variables to construct a large number of multidimensional Gaussian distributions to fit a complex distribution.

As the CVAE can learn a complex probability distribution in the latent space, an unlimited number of random samples can be sampled by the learned distribution. In the probabilistic U-Net, the features of the generated sample are then concatenated to the output of the last layer in U-Net. Finally, three $1 \times 1$ convolutions are applied to combine the concatenated information and map it to the desired number of classes.
U-Net has two characteristics. Firstly, it can model the joint probability of all pixels in the medical image, so the produced segmentation maps can interpret the entire images consistently. Secondly, it models a complex distribution by all training samples, so it also learns the segmentation hypotheses with low probabilities. Due to the above characteristics, the probabilistic U-Net delivers better performance than network ensembles [3], M-heads [5,6], and other related methods.

Figure 3. Schematic diagram of the network structure of PHiSeg and the probabilistic U-Net. Blue boxes represent the feature maps. Arrows represent the flow of operations, where the blue downward-turning arrow contains the downsampling operation and the blue upward-turning arrow contains the upsampling operation. This is a schematic diagram. For more details of the network structures of PHiSeg and the probabilistic U-Net, please refer to [9,10].

2.3. PHiSeg

Christian et al. demonstrate that the probabilistic U-Net has limited diversity when generating the segmentation hypotheses, which may be due to the following two reasons. Firstly, the randomly generated sample is concatenated to the output of the last layer in U-Net, so randomness is only introduced in the highest resolution level of U-Net. Secondly, the randomness is only introduced by channel concatenation, thereby the final three $1 \times 1$ convolutional layers can choose to ignore the random information brought in from the latent space.

Due to the above two reasons, Christian et al. proposed a hierarchical variant of the probabilistic U-Net, called PHiSeg [10]. The network structure of PHiSeg for two resolution levels is shown in Figure 3b. Compared with the probabilistic U-Net shown in Figure 3a, PHiSeg expands the structures of the prior net and the posterior net to the hierarchical multi-resolution structures. The latent space is also learned separately at each downsampled resolution level, and the random segmentation variants are generated from
each latent space and are directly input to the likelihood network instead of being input to U-Net by channel concatenation. The likelihood network outputs multiple segmentation hypotheses, which are then resized to the same size of the ground truth and compared with the ground truth to calculate the loss values. Experimental results show that PHiSeg can generate segmentation hypotheses that closely match the ground-truth distribution.

3. The Proposed HPS-Net

PHiSeg achieved state-of-the-art results on solving the uncertainty of the plausible segmentation hypotheses. However, as another type of uncertainty, the uncertainty of segmentation performance has not drawn enough attention. To solve this problem, in this section, we present the hierarchical predictable segmentation network (HPS-Net), which consists of a new network structure, a new loss function, and a cooperative training mode.

3.1. Network Structure

HPS-Net is a multi-task network. The first task of HPS-Net is to generate realistic and diverse segmentation hypotheses, which is the same as PHiSeg and other related methods. The second task is to provide the measure predictions including precision, accuracy, the true positive rate (TPR), the true negative rate (TNR), or other measures. HPS-Net includes four sub-networks, namely, the posterior network, the prior network, the likelihood network, and the measure network. The posterior network, the prior network, and the likelihood network are for medical image segmentation, the structures of which are similar to PHiSeg [10]. The measure network is for predicting the different measurement values. Figure 4 illustrates the detailed network structure of the proposed HPS-Net for three latent levels and four resolution levels.

![Network structure of the proposed HPS-Net for three latent levels and four resolution levels. HPS-Net includes four sub-networks, namely, the posterior network, the prior network, the likelihood network, and the measure network.](image)

The inputs of the posterior network are the medical image and the corresponding ground-truth segmentation, while the input of the prior network is only the medical image. The goal of these two sub-networks is to learn the latent spaces corresponding to the posterior probability distribution and the prior probability distribution, respectively, and maximize the similarity of the learned latent spaces of these two sub-networks. From Figure 4, we can see that the posterior network and the prior network are sym-
metrical. Such a symmetric network structure allows us to learn the reasonable latent spaces in a confrontational way. With the learned latent spaces, random segmentation variants can be generated, which are then to be input to the likelihood network. With the random segmentation variants as the input, the likelihood network is trained to generate diverse segmentation hypotheses at multiple resolution levels. It should be noted that the resolution levels need at least one more level than the latent levels, and whether each resolution level contains a latent level is optional. The example in Figure 4 has a total of four resolution levels and three latent levels. Figure 5 shows the example of five resolution levels and two latent levels. We obtained the best results in the experiments with seven resolution levels and five latent levels.

![Figure 5. Example of the network structure of five resolution levels and three latent levels.](image)

The final and the most important sub-network is the measure network, which enables HPS-Net to generate the predicted measurement values. The measure network takes the medical image and its corresponding segmentation hypotheses as the input. To fuse the information, the segmentation hypotheses at different levels are upsampled to the same size of the medical image and then fused with the medical image by channel concatenation. The resulting multi-channel image is then input into the backbone network. Here, we used the SE-Inception-V4 (the code and the details of SE-Inception-V4 can be found in https://github.com/taki0112/SENet-Tensorflow, accessed on 27 October 2021) as the backbone network, which is a combination of Inception-v4 [17] and the squeeze-and-excitation (SE) block [18]. As a convolutional neural network, Inception-v4 has an excellent performance in the field of image recognition. The SE block can adaptively adjust the relationship between different channels by explicitly modeling interdependencies between channels. As the input of the measure network is a multi-channel image and the importance of each channel is quite different, applying the SE block to Inception-v4 can further improve the performance at a slight additional computational cost. After the combination, SE-Inception-V4 processes the image by convolution, pooling, residual connection, squeeze, excitation, and other operations, breaking down the image into features. The result of this process feeds into the final fully connected layer that drives the final predicted measurement value.

### 3.2. Loss Functions

In the posterior network, we used the Kullback–Leibler divergence to penalize the difference between the posterior distribution and the prior distribution. For the likelihood network, cross-entropy loss was used for each latent level to penalize the difference between the segmentation hypothesis and the ground truth of the segmentation (mask). Following the practice of PHiSeg [10], if an input medical image had multiple corresponding masks, then one mask was randomly selected for computing. As all masks were plausible, random selection was therefore reasonable and efficient in this case.
We proposed measure loss to penalize the difference between the predicted measurement value and the ground truth of the measurement. The input of the measure loss includes the ground truth segmentations, the segmentation hypothesis output from each latent level, and the predicted measurement value. The measure loss is formulated as follows:

\[
\mathcal{L}_{ms} = \frac{1}{l} \sum_{i=1}^{l} \mathcal{L}_i, \tag{1}
\]

\[
\mathcal{L}_i = \begin{cases} 
\frac{1}{N} \sum_{n=1}^{N} (M_n - U_n)^2, & M_n > U_n \\
\frac{1}{N} \sum_{n=1}^{N} (M_n - L_n)^2, & M_n < L_n \\
0, & L_n \leq M_n \leq U_n
\end{cases}, \tag{2}
\]

\[
U_n = \max \{M(S_n, S_{nj}) : j = 1, 2, ..., m\}, \tag{3}
\]

\[
L_n = \min \{M(S_n, S_{nj}) : j = 1, 2, ..., m\}, \tag{4}
\]

where \( \mathcal{L}_{ms} \) represents the measure loss, \( i \) is the number of latent levels, \( \mathcal{L}_i \) represents the sub-loss of the \( i \)th latent level, \( N \) is the batch size, \( M_n \) is the predicted measurement value of the \( n \)th sample, \( U_n \) represents the upper bound of the ground truth measurement value of the \( n \)th segmentation hypothesis, \( L_n \) represents the lower bound of the ground truth measurement value of the \( n \)th segmentation hypothesis, \( S_n \) denotes the \( n \)th segmentation hypothesis, \( m \) is the number of ground truth, \( S_{nj} \) denotes the \( j \)th ground truth of the \( n \)th segmentation hypothesis, and \( M(\cdot, \cdot) \) calculates the measurement value, which can be TPR, TNR, precision, and others.

3.3. Training Procedure

This section describes our cooperative training mode. The training process mainly includes two parts—forward propagation and backward propagation. In forward propagation, each batch of the medical images is firstly input into the posterior network and the prior network. The outputs of the two sub-networks represent multiple multidimensional normal distributions, so as to fit their learned latent spaces respectively. Then, the Kullback–Leibler divergence of each level is calculated as the loss value based on the outputs of the two sub-networks to describe the difference between the learned latent spaces. With the learned latent space of the prior network, samples can be generated and then be input into the likelihood network to get the segmentation hypotheses. By comparing the output segmentation hypotheses and the ground-truth segmentations, the cross entropy losses are computed. After that, the segmentation hypotheses are input into the measure network to generate the predicted measurement value, where the measure loss is applied for calculating the difference between the predicted measurement value and the ground-truth measurement value. The ground-truth measurement value is initially unknown, but it can be calculated with the output segmentation hypotheses and the ground-truth segmentations.

In backward propagation, the gradients of the loss function with respect to the weights of the network are calculated, and this calculation process proceeds backward through the network from the final layer to the first layer. Specifically, the network parameters in the posterior network and the prior network are updated by the calculated Kullback–Leibler divergence of each level. The network parameters in the likelihood network are updated by the cross entropy losses, and the measure loss corresponds to the parameter update in the measure network. It should be noted that although there are connections between the likelihood network and the measure network, we limited the backpropagation starting from the measure loss to the range of the measurement network, so as to avoid interference with the prediction of the segmentation hypothesis. The model was trained using the
Adam optimizer with a learning rate of $10^{-3}$ and a batch-size of 12. All the input images and the ground-truth segmentations were resized to $128 \times 128$. Batch-normalization was applied on all non-output layers. All models were trained for 36 h on an NVIDIA TESLA V100 GPU. After a model is trained, the segmentation hypotheses and the corresponding measurement values for an input image can be generated by first inputting the image to the prior network, then obtaining the samples using the prior network and finally generating the segmentation hypotheses and the measurement values using the likelihood network and the measure network, respectively.

4. Experiments and Results

In the experiments, two aspects of HPS-Net were evaluated. The first aspect was to evaluate whether HPS-Net can perform as well as PHiSeg after introducing the measure network. The second aspect was to evaluate the performance of HSP-Net on predicting different measurement values.

Experiments were conducted on the LIDC-IDRI dataset [11] and the ISIC2018 dataset [12]. The LIDC-IDRI dataset contains 1018 thoracic CT images from 1010 subjects with lesions annotated by four experienced thoracic radiologists. The original in-plane resolution of the CT images was between 0.461 mm and 0.977 mm. In the experiments, all CT images were normalized to $0.5 \text{ mm} \times 0.5 \text{ mm}$ resolution. Based on the lesion positions annotated by the radiologists, the lesion patches of size $128 \times 128$ pixels were cropped out. The processed dataset was then partitioned into the ratio of 60:20:20 for training, testing, and validation, respectively. The ISIC2018 dataset consists of 3694 skin lesion images, including a training set, a validation set, and a test set. The training set contains 100 images for training, 100 images for validation, and 1000 images for testing, respectively. The initial size of images in ISIC2018 varied from $1024 \times 768$ to $512 \times 384$. In our experiments, these images were normalized to $512 \times 512$. To compare with PHiSeg [10], we followed the setting of [10] in all experiments, setting seven resolution levels for both PHiSeg and HPS-Net.

In the first experiment, we firstly trained the models with the masks from all four radiologists and then trained the models with only one mask from a single radiologist. When training with all masks, one mask was randomly selected per image in each batch. When training with one mask, only the mask from the same radiologist was used for the corresponding image in each batch. We used a conventional Dice score to gauge the similarity between an output segmentation hypothesis and a mask. To compare a set of segmentation hypotheses and a set of masks, namely, to compare two sets of segmentation distributions, we used the generalized energy distance [9,19]:

$$D_{GED}(P_{gt}, P_{\text{out}}) = 2\mathbb{E}[d(S, Y)] - \mathbb{E}[d(S, S')] - \mathbb{E}[d(Y, Y')]$$

where $Y$ and $Y'$ are independent masks from the ground-truth distribution $P_{gt}$, $S$ and $S'$ are independent segmentations from the predicted distribution $P_{\text{out}}$, and $d(\cdot, \cdot) = 1 - \text{IoU}(\cdot, \cdot)$. $1 - \text{IoU}(\cdot, \cdot)$ has been proved to be a metric [20,21]. $D_{GED}$ is also a metric in this case if $d(\cdot, \cdot)$ is also a metric, which has been proved in [22]. Like [10], we also used the average normalized cross correlation (NCC) to evaluate the correlation between the predicted distribution and the ground-truth distribution from a visual perspective:

$$S_{NCC}(P_{gt}, P_{\text{out}}) = \mathbb{E}_{S \sim P_{gt}}[\text{NCC}(\mathbb{E}_{S \sim P_{\text{out}}}[CE(\mathbb{E}_{S \sim P_{\text{out}}}[CE(\mathbb{E}_{S \sim P_{\text{out}}}(Y, S)))]))$$

where $CE$ denotes cross entropy and $\mathbb{E}$ is the mean segmentation prediction.

Table 1 show the results of the first experiment, which compares the segmentation performance of different methods on the LIDC-IDRI dataset. From Table 1, we can observe that, using all annotations, HPS-Net ($L = 5$) and PHiSeg ($L = 5$) performed better than prob. U-Net in terms of $D_{GED}$ and $S_{NCC}$. With single annotation, HPS-Net ($L = 5$) and PHiSeg ($L = 5$) outperformed prob. U-Net and det. U-Net in terms of $D_{ICC}$, $D_{GED}$, and $S_{NCC}$. Additionally, we can observe that HPS-Net showed slightly better performance than
PHiSeg with the same number of resolution levels in the cases of both single annotation and all annotations, but the performance difference was small. This is in line with our expectation, as we limited the backpropagation starting from the measure loss to the range of the measure network; the prediction performance of the likelihood network was therefore not affected by the measure network.

In the second experiment, we trained the HPS-Net models with different latent levels and different numbers of radiologists to evaluate the performance of HSP-Net on predicting different measurement values. Table 2 shows the results of the second experiment, from which we could observe the following:

- Compared with HPS-Net (L = 1) and HPS-Net (L = 3), HPS-Net (L = 5) had the lowest mean squared error (MSE) and the lowest standard deviation (std.) in most cases.
- The predictions on TNR were much more accurate than the predictions on TPR and precision, where the MSE ± std. on TPR was lower than the MSE ± std. on precision.
- The MSE ± std. on TNR was as low as 0.0001 ± 0.0062 (all annotation) and 0.0010 ± 0.0219 (single annotation), which is close to the requirements of practical application. However, the MSE ± std. on TPR and precision was still as high as 0.0938 ± 0.1080 and 0.1160 ± 0.1449 with single annotation and L = 5, which is far from practical application and still needs to be improved.

Table 1. Segmentation performance of different methods on LIDC-IDRI dataset.

| # Radiologists | \(D^2_{GED}\) | \(S_{NCC}\) | Dice |
|----------------|---------------|-------------|------|
| Prob. U-Net    | 0.2393        | 0.7749      | -    |
| PHiSeg (L = 1) | 0.2934        | 0.7944      | -    |
| PHiSeg (L = 5) | 0.2248        | 0.8453      | -    |
| HPS-Net (L = 1)| 0.2410        | 0.8025      | -    |
| HPS-Net (L = 5)| 0.2218        | 0.8414      | -    |
| Det. U-Net     | 0.2025 ± 0.2329 | 0.1441 ± 0.1450 | 0.3671 ± 0.3736 |
| HPS-Net (L = 3)| 0.1536 ± 0.1940 | 0.0019 ± 0.0254 | 0.3179 ± 0.3442 |
| HPS-Net (L = 5)| 0.0953 ± 0.1109 | 0.0001 ± 0.0062 | 0.1123 ± 0.1696 |

Table 2. Mean squared error (MSE) and standard deviation (std.) of squared error of different methods when predicting TPR, TNR, and precision on LIDC-IDRI dataset.

| # Radiologists | MSE ± Std. of TPR | MSE ± Std. of TNR | MSE ± Std. of Precision |
|----------------|-------------------|-------------------|-------------------------|
| HPS-Net (L = 1)| 0.1096 ± 0.1788   | 0.0019 ± 0.0302   | 0.4248 ± 0.4298         |
| HPS-Net (L = 3)| 0.1669 ± 0.2181   | 0.0009 ± 0.0185   | 0.3179 ± 0.3719         |
| HPS-Net (L = 5)| 0.0938 ± 0.1080   | 0.0010 ± 0.0219   | 0.1160 ± 0.1449         |

To further verify the performance of HPS-Net on image segmentation and measure prediction, an additional experiment was conducted on the ISIC2018 dataset [12], where we trained the HPS-Net models with different latent levels and compared it with three benchmark methods, namely, det. U-Net, prob. U-Net, and PHiSeg. From Table 3, we can observe that HPS-Net had the highest Dice score, which means it had the best segmentation performance. Similarly, the MSE ± std. of TNR was the lowest among all three measure predictions, while the MSE ± std. of precision was still the highest.
Table 3. Quantitative results of different methods on ISIC2018 dataset.

| Method                | Dice     | MSE ± Std. of TPR | MSE ± Std. of TNR | MSE ± Std. of Precision |
|-----------------------|----------|-------------------|-------------------|-------------------------|
| Det. U-Net            | 0.6359   | -                 | -                 | -                       |
| Prob. U-Net           | 0.6312   | -                 | -                 | -                       |
| PHiSeg (L = 1)        | 0.6939   | -                 | -                 | -                       |
| PHiSeg (L = 5)        | 0.7082   | -                 | -                 | -                       |
| HPS-Net (L = 1)       | 0.7381   | 0.0487 ± 0.2153   | 0.0162 ± 0.0082   | 0.1856 ± 0.2632         |
| HPS-Net (L = 5)       | 0.7527   | 0.0401 ± 0.1934   | 0.0135 ± 0.0074   | 0.1498 ± 0.2315         |

5. Conclusions

In this study, we proposed the hierarchical predictable segmentation network (HPS-Net) for medical image segmentation. HPS-Net utilizes our proposed measure network to obtain the ability of predicting the segmentation reliability and utilizes our proposed new loss function for computing the difference between the true measurement value and the predicted measurement value. As a multi-task network, HPS-Net generates diverse segmentation hypotheses and provides effective measure predictions for corresponding segmentation hypotheses. From another perspective, HPS-Net can also provide the segmentation hypothesis based on a specified measurement value. The efficiency of the proposed method was demonstrated on the LIDC-IDRI dataset and the ISIC2018 dataset. Through experiments, we confirmed that it is possible to predict the reliability of the segmentation results. The HPS-Net can contribute to a certain extent in clinical diagnoses to prevent misdiagnosis and help radiologists to make better treatment plans.

Author Contributions: Conceptualization, X.W. and H.W.; writing—original draft, X.W.; writing—review and editing, F.Y. and W.M. All authors read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 6210021948, 62076117) and the Jiangxi Key Laboratory of Smart Cities (Grant No. 20192BCD40002).

Data Availability Statement: The data presented in this study are available on request from the first author.

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

References

1. Jiang, J.; Fares, A.; Zhong, S.H. A context-supported deep learning framework for multimodal brain imaging classification. IEEE Trans. Hum.-Mach. Syst. 2019, 49, 611–622. [CrossRef]
2. Warfield, S.K.; Zou, K.H.; Wells, W.M. Validation of image segmentation and expert quality with an expectation-maximization algorithm. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Berlin, Heidelberg, 25–28 September 2002; Springer: Berlin/Heidelberg, Germany, 2002; pp. 298–306.
3. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems; Curran Associates Inc.: Long Beach, CA, USA, 2016; pp. 6402–6413.
4. Guzman-Rivera, A.; Batra, D.; Kohli, P. Multiple choice learning: Learning to produce multiple structured outputs. In Advances in Neural Information Processing Systems; Curran Associates Inc.: Lake Tahoe, NV, USA, 2012; pp. 1799–1807.
5. Rupprecht, C.; Laina, I.; DiPietro, R.; Baust, M.; Tombari, F.; Navab, N.; Hager, G.D. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3591–3600.
6. Ilg, E.; Çiçek, Ö; Galesso, S.; Klein, A.; Makansi, O.; Hutter, F.; Brox, T. Uncertainty estimates for optical flow with multi-hypotheses networks. arXiv 2018, arXiv:1802.07095.
7. Sohn, K.; Lee, H.; Yan, X. Learning structured output representation using deep conditional generative models. In Advances in Neural Information Processing Systems; Curran Associates Inc.: Montreal, QC, Canada, 2015; pp. 3483–3491.
8. 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.
9. Kohl, S.A.; Romera-Paredes, B.; Meyer, C.; De Fauw, J.; Leddam, J.R.; Maier-Hein, K.H.; Eslami, S.M.; Rezende, D.J.; Ronneberger, O. A probabilistic u-net for segmentation of ambiguous images. arXiv 2018, arXiv:1806.05034.
10. Baumgartner, C.F.; Tezcan, K.C.; Chaitanya, K.; Hötker, A.M.; Muehlematter, U.J.; Schawkat, K.; Becker, A.S.; Donati, O.; Konukoglu, E. Phiseg: Capturing uncertainty in medical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 119–127.

11. Armato, S.G., III; McLennan, G.; Bidaut, L.; McNitt-Gray, M.F.; Meyer, C.R.; Reeves, A.P.; Zhao, B.; Aberle, D.R.; Henschke, C.I.; Hoffman, E.A.; et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. *Med Phys.* 2011, 38, 915–931. [CrossRef] [PubMed]

12. Codella, N.C.; Gutman, D.; Celebi, M.E.; Helba, B.; Marchetti, M.A.; Dusza, S.W.; Kalloo, A.; Liopyris, K.; Mishra, N.; Kittler, H.; et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; pp. 168–172.

13. Fan, D.P.; Ji, G.P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. Pranet: Parallel reverse attention network for polyp segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 263–273.

14. Feng, S.; Zhao, H.; Shi, F.; Cheng, X.; Wang, M.; Ma, Y.; Xiang, D.; Zhu, W.; Chen, X. CPFNet: Context pyramid fusion network for medical image segmentation. *IEEE Trans. Med Imaging* 2020, 39, 3008–3018. [CrossRef] [PubMed]

15. Sun, J.; Darbehani, F.; Zaidi, M.; Wang, B. Saunet: Shape attentive u-net for interpretable medical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 797–806.

16. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.

17. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.

18. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.

19. Székely, G.J.; Rizzo, M.L. Energy statistics: A class of statistics based on distances. *J. Stat. Plan. And Inference* 2013, 143, 1249–1272. [CrossRef]

20. Kosub, S. A note on the triangle inequality for the Jaccard distance *Pattern Recognit. Lett.* 2019, 120, 36–38. [CrossRef]

21. Lipkus, A.H. A proof of the triangle inequality for the Tanimoto distance. *J. Math. Chem.* 1999, 26, 263–265. [CrossRef]

22. Klebanov, L.B.; Beneš, V.; Saxl, I. *N-Distances and Their Applications*; Charles University in Prague, the Karolinum Press: Prague, Czech Republic, 2005.