Small Lesion Segmentation in Brain MRIs with Subpixel Embedding

Alex Wong[1(✉)], Allison Chen[1], Yangchao Wu[1], Safa Cicek[1], Alexandre Tiard[1], Byung-Woo Hong[2], and Stefano Soatto[1]

[1] University of California, Los Angeles, CA, USA
`alexw@cs.ucla.edu`
[2] Chung-Ang University, Seoul, Korea

Abstract. We present a method to segment MRI scans of the human brain into ischemic stroke lesion and normal tissues. We propose a neural network architecture in the form of a standard encoder-decoder where predictions are guided by a spatial expansion embedding network. Our embedding network learns features that can resolve detailed structures in the brain without the need for high-resolution training images, which are often unavailable and expensive to acquire. Alternatively, the encoder-decoder learns global structures by means of striding and max pooling. Our embedding network complements the encoder-decoder architecture by guiding the decoder with fine-grained details lost to spatial down-sampling during the encoder stage. Unlike previous works, our decoder outputs at 2× the input resolution, where a single pixel in the input resolution is predicted by four neighboring subpixels in our output. To obtain the output at the original scale, we propose a learnable down-sampler (as opposed to hand-crafted ones e.g. bilinear) that combines subpixel predictions. Our approach improves the baseline architecture by ≈11.7% and achieves the state of the art on the ATLAS public benchmark dataset with a smaller memory footprint and faster runtime than the best competing method. Our source code has been made available at: https://github.com/alexklwong/subpixel-embedding-segmentation.

1 Introduction

A stroke occurs when a lack of blood flow prevents brain tissue from receiving adequate oxygen and nutrients. This condition affects over 795,000 people annually [28]. The severity of the outcome, including disability and paralysis, depends on the location and intensity of the stroke, as well as the time of diagnosis [2,30]. Preserving cognitive and motor functions, therefore, hinges on localizing stroke

---

A. Wong and A. Chen—Authors with equal contributions.

lesions quickly and precisely. However, doing so manually requires expert knowledge, is time consuming, and is ultimately subjective [11,13].

We focus on automatically segmenting ischemic stroke lesions, which account for 87% of all strokes [28], from T1-weighted anatomical magnetic resonance imaging (MRI) brain scans. These lesions are characterized by high variability in location, shape, and size – the latter two are problematic for conventional convolutional neural networks (CNNs) where precision of irregularly shaped lesion boundaries and recall of small lesions are critical measures of success. Due to aggressive spatial downsampling (i.e. max pooling, strided convolutions) customary in CNNs, details of local structures are lost in the process. Yet, the spatial downsampling is necessary for obtaining a global representation of the input while using fixed-size filters with limited receptive fields. The outcome of which are segmentations with ambiguous boundaries between lesion and normal tissues and missed lesions that occupy small number of voxels in the MRIs.

We propose to retain small local structures by learning an embedding that maps the input to high dimensional feature maps of *twice* the input resolution. Unlike the typical CNN, we do not perform lossy downsampling on this representation; hence, the embedding preserves local structures, but lacks global context. When combined with the standard encoder-decoder e.g. U-Net [19], the embedding complements the encoder-decoder by supplying the decoder with fine-grained detail information to guide segmentation. Our network also outputs at twice the resolution of the input, representing each element in the input with a 2×2 neighborhood of predictions. The final output is obtained by combining the four predictions (akin to an ensemble) as a weighted sum where the contribution of each prediction is learned from the data. Our design not only enables the network to produce robust segmentations but also localize small lesions (Fig. 3).

Our contributions include (i) an embedding function that preserves fine-grained details of the input by mapping it to larger spatial dimensions, (ii) a neural network architecture that leverages the complementary strengths of the proposed embedding and an encoder-decoder to produce predictions at twice the input resolution, and (iii) a learnable downsampler that combines local predictions in an ensemble fashion to yield robust segmentations at the input resolution. Our approach improves the baseline U-Net architecture by $\approx 11.7\%$ and achieves the state of the art on the ATLAS [11,12] dataset with lower computational burden than the best competing method.

2   Related Work

Lesion Segmentation. Early works [4] aggregated classification results for the center pixel of patches sampled from an image. However, [4] lacked global context, so [21] addressed this with multi-stage cascaded hierarchical models. More recent works build upon the U-Net [19], a 2D fully-convolutional network with skip connections and up-convolutions. For example, [14] used a Dual Path Network [3] encoder while [26] leveraged dilated convolutions to inexpensively increase receptive fields. Furthermore, [1] fused the U-net with other

high-performing modules, the BConvLSTM [24] and the SENet [8], and [18] introduced X-blocks to the U-Net, leveraging depthwise separable convolutions to reduce computational load. [31] used skip connections between successive encoder resolutions to prevent the loss of features and ConvLSTM [23] modules to maintain localization.

Recent works also leveraged 3D architectural backbones to improve localization. [32] performed 3D convolutions on a subsection of the scan and fused the results with 2D convolutions. [9] proposed an attention gate to combine 2D segmentations along the axial, sagittal, coronal planes into a 3D volume. However, these works use significantly larger memory footprints and 3D convolutions are computationally expensive – limiting the models' practicality. We note that while conventional architectures perform well globally (i.e. recovering the coarse shape of lesions) they struggle to segment small lesions that blend into the background.

Super-Resolution. There is an abundance of works in natural images super-resolution [5, 6, 22, 25, 29] and a growing number in medical imaging. [20] proposed to map MRI images from low to high-resolution with an overcomplete dictionary. [16] leveraged SRCNN [5] for super-resolving 2D MRI images and fused them to obtain a 3D volume. [17] handled arbitrary scaling factors with a 3D architecture for multi-modal 3D data. However, these works require low and high-resolution image pairs for training and are limited to the super-resolution task while our method does not rely on a larger resolution ground truth. More recently, [27] introduced Kite-Net, an upsampling encoder that outputs a latent at $8\times$ resolution followed by a max-pooling decoder to downsample back to the original resolution. Kite-Net is used in parallel with a U-Net for lesion segmentation. Our approach draws inspiration from super resolution and latent over-representations as methods to retain local structure that are often lost in spatial downsampling. However, unlike [27], we avoid downsampling the latent with pooling (which discards information), and instead employ lossless space-to-depth and depth-to-space [22] operations to retain fine-grained details. Furthermore, we propose to learn a subpixel embedding at $2\times$ the original resolution to guide our segmentation, which uses a much smaller memory footprint than [27]. We show that our approach can capture small lesions that are missed by [18, 19, 27, 31, 32].

3   Method

We propose a method to partition a 3D MRI volume $X \in \mathbb{R}^{C \times H \times W}$ into lesion (positive, 1) and normal (negative, 0) classes. Our method takes, as input, a 3D slice of c consecutive 2D images $x \in \mathbb{R}^{c \times H \times W}$ (c is an odd integer) from X and predicts the binary segmentation for the image $\bar{x} \in \mathbb{R}^{1 \times H \times W}$, the $\frac{c+1}{2}$-th image of x. In other words, x is a sliding window of c images centered at a target image $\bar{x}$. To avoid sampling out of bounds, we perform mean padding of size $\frac{c-1}{2} \times H \times W$ on both sides of X before sampling x (see Sec. 1 of Supp. Mat. for more details). To segment a single image $\bar{x}$, we propose to learn a deep
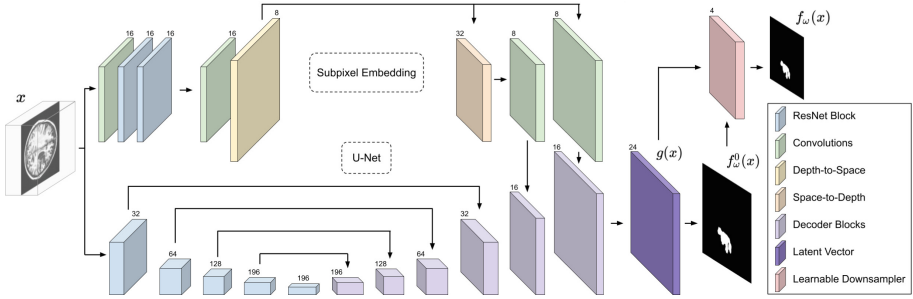
Fig. 1. *Network architecture.* SPiNis comprised of (i) a U-Net based encoder-decoder that produces subpixel predictions $f_\omega^0(x)$ at $2\times$ the input resolution, which are guided by (ii) a subpixel embedding that captures local structure. The final output $f_\omega(x)$ is achieved by combining local predictions in a 2×2 neighborhood as a weighted sum based on the per element contribution predicted by a (iii) learnable downsampler.

neural network f_ω, parameterized by ω, where $f : \mathbb{R}^{c \times H \times W} \mapsto [0, 1]^{1 \times H \times W}$ is a function that takes the 3D slice x as an input and outputs the sigmoid response $f_\omega(x)$, a confidence map corresponding to lesions in $\bar{x}$. To obtain the binary segmentation of X, we aggregate our predictions by running f_ω for all x and setting any response greater than a threshold of 0.5 to the lesion class. We note that our method can be extended to multi-class segmentation simply by expanding our output to $[0, 1]^{K \times H \times W}$ for K classes, and choosing the class with highest response, i.e. $\arg\max f_\omega(\cdot)$, to yield the segmentation.

3.1   Network Architecture

Our network f_ω (Fig. 1) is composed of two modules: (i) an encoder-decoder (based on U-Net [19]) that outputs at $2\times$ the input resolution, e.g. $2H \times 2W$, whose predictions are guided by (ii) a network that maps the input x to a high dimensional embedding space also at twice the input resolution. The result is a confidence map comprised of "subpixel" predictions – the output class for each input pixel is represented by four predictions within a 2×2 neighborhood. Rather than using hand-crafted downsampling techniques (e.g. bilinear, nearest neighbor) to obtain the output at the original ($1\times$) spatial resolution, we propose a learnable downsampler that predicts the weight, or contribution, of each subpixel prediction in a local region corresponding to the pixel in the $1\times$ resolution. For simplicity, we refer to our embedding function as a subpixel embedding and our overall architecture (f_ω) as a subpixel network or "SPiN" for short (Fig. 1).
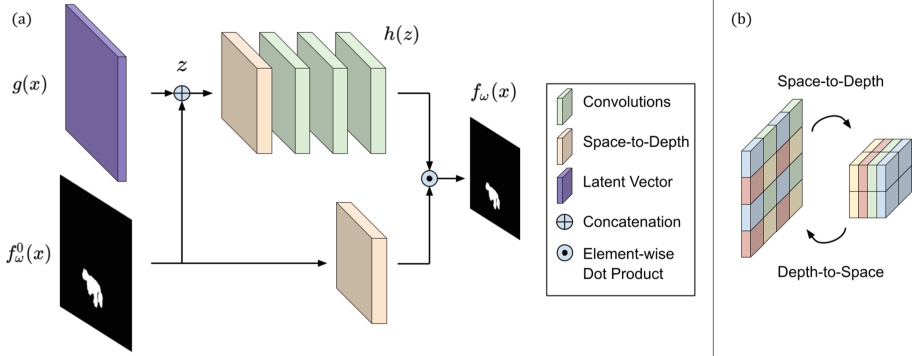
Fig. 2. *Learnable Downsampler, Space-to-Depth and Depth-to-Space.* (a): Learnable Downsampler predicts the contribution $h(z)$ of each subpixel prediction in $f_\omega^0(x)$ by conditioning on $f_\omega^0(x)$ and the latent vector $g(x)$. Subpixel predictions $f_\omega^0(x)$ are rearranged to the resolution of the input using Space-to-Depth. The final output $f_\omega(x)$ is produced by taking the element-wise dot product between $h(z)$ and the reshaped $f_\omega^0(x)$. (b) Space-to-Depth reduces resolution by rearranging elements from the spatial dimensions into the channel dimensions, where each 2×2 neighborhood is reshaped to a 4 element vector. Depth-to-Space conversely performs spatial expansion by rearranging elements from the channel dimensions to height and width dimensions.

Subpixel embedding consists of feature extraction and spatial expansion phases. Feature extraction is performed by two ResNet blocks [7] with 16 filters per layer; we also use stride of 1 and zero-padded edges to minimize spatial reduction. The extracted $16 \times H \times W$ feature maps are fed to a depth-to-space module [22] that rearranges elements from the channel dimension to the height and width dimensions (see Fig. 2-(b)). The resulting set of $4 \times 2H \times 2W$ feature maps with twice the spatial resolution then undergoes a 1×1 and a 3×3 convolution layers, with 8 filters each. The resulting $8 \times 2H \times 2W$ high dimensional feature maps, produced by our subpixel embedding function, resolve fine local details by increasing the feature map resolution and thus representing information at each pixel location with four "subpixel" feature vectors.

When used as skip connections, these embeddings complement the standard U-Net architecture that obtains a global representation of the input by spatial downsampling (striding and max pooling), which naturally discards local detail. Hence, we propose to inject these embeddings into the decoder via feature concatenation at the original (1×) resolution and at the 2× output resolution. To reduce the height and width dimensions of the embeddings to match the feature maps at the 1× resolution, we propose a space-to-depth module, which performs the inverse operation of depth-to-space (see Fig. 2-(b)), yielding $32 \times H \times W$ feature maps. Unlike striding and pooling, the depth-to-space operation is information preserving as it rearranges feature vectors from the height and width dimensions to their channel dimension. The result is fed through a 3×3 convolutional layer with 8 filters and concatenated with the feature maps of the decoder at the 1× resolution. Similarly, the embeddings at 2× resolution undergo

a separate 3×3 convolution to yield the output resolution guidance before being concatenated with their corresponding feature maps in the decoder. Finally, the $2\times$ decoder output $f_\omega^0(x) \in [0,1]^{1 \times 2H \times 2W}$ is produced by convolving a single 3×3 filter over the resulting latent vector $g(x) \in \mathbb{R}^{24 \times 2H \times 2W}$. We use subpixel guidance (SPG) to refer to the process of learning and injecting the embedding as skip connections, which substantially helps with localizing small lesions missed by previous works [18,19,31,32] (see Fig. 3). We note that SPG is light-weight and only uses 16K parameters.

Learnable downsampler takes the concatenation $z = [g(x); f_\omega^0(x)]$ of the latent vector $g(x)$ and the $2\times$ resolution output $f_\omega^0(x)$ and predicts $h(z)$, where $h : \mathbb{R}^{25 \times 2H \times 2W} \mapsto [0,1]^{4 \times H \times W}$. In other words, $h(z)$ is a set of $4 \times H \times W$ values that determine the contribution of each subpixel prediction in a 2×2 neighborhood of $f_\omega^0(x)$. To achieve this, we first perform space-to-depth on z to rearrange each 2×2 neighborhood into a 4 element vector. This is followed by two 3×3 convolutions of 16 filters and a 1×1 convolution with 4 filters. $h(z)$ is the softmax response of the result along the channel dimension.

To obtain the final output $f_\omega(x)$, we utilize space-to-depth to rearrange $f_\omega^0(x)$ into the shape of $4 \times H \times W$ (to match the shape of $h(z)$) and take its element-wise dot product with $h(z)$. With an abuse of notation, $f_\omega(x) = f_\omega^0(x) \cdot h(z)$. Because $h(z)$ is conditioned on the latent vector $g(x)$ of the input, the predicted weights respect lesion boundaries to yield detailed segmentations. This is unlike bilinear or nearest-neighbor downsampling where weights are predetermined and independent of the input. We note that our learnable downsampler is also lightweight and only consists of 11K parameters.

3.2 Loss Function

We assume a training set of $\{(x^{(n)}, \bar{y}^{(n)})\}_{n=1}^N$, where $\bar{y}^{(n)}$ is the ground truth corresponding to $\bar{x}^{(n)}$, the image located at the center of $x^{(n)}$. To train SPiN, we minimize the standard binary cross entropy loss,

$$\ell(y, \bar{y}) = \frac{1}{|\Omega|} \sum_{u \in \Omega} -\big(\bar{y}(u) \log y(u) + (1 - \bar{y}(u)) \log(1 - y(u))\big), \qquad (1)$$

where $\Omega \subset \mathbb{R}^2$ denotes the spatial image domain, u a pixel coordinate, and $y = f_\omega(x)$ the network output. The loss over the training set of N samples reads

$$L(\omega) = \frac{1}{N} \sum_{n=1}^N \ell(f_\omega(x^{(n)}), \bar{y}^{(n)})). \qquad (2)$$

We note that previous works [31,32] used soft Dice loss (an approximation of the true Dice score) to counter the class imbalance between normal and lesion tissues, characteristic in the lesion segmentation problem. However, a minimizer of cross entropy equivalently minimizes Dice, and empirically, we found that directly minimizing cross entropy yields better performance for our model. We

Table 1. *Evaluation metrics.* IOU denotes Intersection Over Union, and DSC denotes Dice similarity coefficient. TP, FN and FP correspond to true positive, false negative and false positive respectively.

Metric	IOU	DSC	Precision	Recall
Definition	$\frac{\text{TP}}{\text{TP+FN+FP}}$	$\frac{2\times\text{TP}}{2\times\text{TP+FN+FP}}$	$\frac{\text{TP}}{\text{TP+FP}}$	$\frac{\text{TP}}{\text{TP+FN}}$

hypothesize that our SPG allows small lesions to be recovered more easily, making our method more conducive to minimizing cross entropy, which is not prone to the noisy training signal inherent in soft Dice. We demonstrate this in row 7 of Table 4 in our ablation studies. Also, we note that our loss can be easily extended for multi-class classification to accommodate multiple lesion categories.

4   Experiments and Results

We demonstrate our method on the Anatomical Tracings of Lesion After Stroke (ATLAS) MRI dataset [11,12], using the metrics defined in Table 1. ATLAS contains 304 T1-weighted MRI scans of stroke patients with corresponding lesion annotations. The data is collected from 11 research sites worldwide, manually annotated, and post-processed (i.e. smoothing and defacing for privacy), leaving 239 patient scans with 189 2D images (197×233 resolution) each. Since no official data split is provided by [11], previous works [18,31,32] evaluated their methods using k-fold cross validation and randomly sampled data splits. However, the value of k and samples within each split varied across works. Due to the lack of consistency, the reported results are not directly comparable. Thus, we propose a training (212 patients) and a held-out testing (27 patients) split to standardize the evaluation protocol for more rigorous comparisons. We provide quantitative comparisons against [18,19,27,31,32] on the proposed training and testing split in Table 2. We also show qualitative (Fig. 3) and quantitative (Table 3) comparisons on segmenting small lesions using a subset of test set: 490 images containing *only* lesions smaller than 100 pixels (0.2% of the image). All reported results for previous works are obtained using their training procedures and open-sourced code. We also provide details on our training and testing split in Sec. 2 of Supp. Mat. and further k-fold cross validation comparisons in Sec. 3 of Supp. Mat.

Implementation Details. Our model is implemented in PyTorch [15] and optimized using Adam [10]. We used an initial learning rate of 3×10^{-4}, decreased it to 1×10^{-4} after 400 epochs, and to 5×10^{-5} after 1400 epochs for a total of 1600 epochs. We choose $c = 5$ for the number images in the input x. During training, $\bar{x}$ and its corresponding x are randomly sampled from X. Training takes $\approx 8\,$h on an Nvidia GTX 1080 GPU, and inference takes ≈ 11 ms per 2D image. For data augmentation, we randomly perform (i) horizontal and vertical flips, (ii) rotation between $-30°$ and $30°$, and (iii) add zero-mean Gaussian noise with
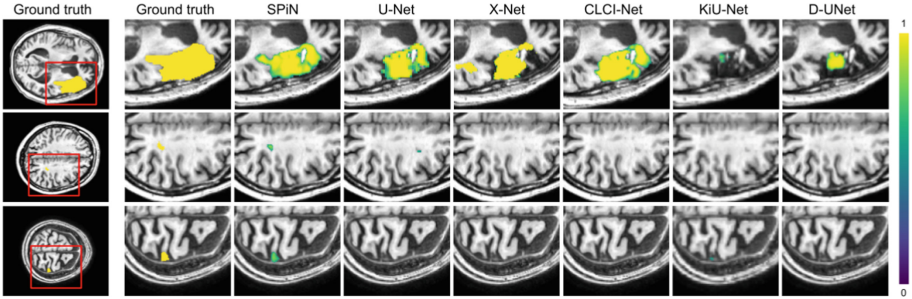
Fig. 3. *Qualitative results on ATLAS.* Columns 2–8 show (zoomed in) head-to-head comparisons across all methods for highlighted areas in column 1. Row 1 demonstrates that SPiNoutperforms existing works in capturing shape and boundary details in medium-sized, irregularly-shaped lesions. Furthermore, rows 2 and 3 demonstrate SPiN's ability to localize small lesions that are missed by other models.

Table 2. *Quantitative comparison on ATLAS.* SPiNoutperforms all methods across all performance metrics. It is also one of the least computationally expensive models, i.e. smallest test time memory footprint, second in training memory usage, and third fastest in runtime per patient (189 images).

Method	Performance metrics				Runtime (s)	Memory usage (GB)	
	DSC	IOU	Precision	Recall		Train	Test
U-Net [19]	0.584	0.432	0.674	0.558	1.375	2.291	1.181
D-UNet [32]	0.548	0.404	0.652	0.521	3.425	15.426	15.426
CLCI-Net [31]	0.599	0.469	0.741	0.536	8.860	7.853	7.853
KiU-Net [27]	0.524	0.387	0.703	0.459	1.05	23.566	1.555
X-Net [18]	0.639	0.495	0.746	0.588	5.046	11.839	11.839
SPiN(Ours)	**0.703**	**0.556**	**0.806**	**0.654**	2.145	3.273	0.803

standard deviation of 1×10^{-2} to training samples. We perform augmentation with a probability of 1 for 1400 epochs and decrease it to 0.5 thereafter so training samples will be closer to the true distribution of the dataset.

ATLAS Test Set. Table 2 shows that our approach outperforms competing methods [18,19,27,31,32] across all evaluation metrics. Specifically, we beat the best performing method X-Net [18] by an average of ≈10.4% with a 72.3% reduction in training memory and a 57.5% runtime reduction during inference. Our approach also uses a smaller memory footprint, containing only ≈5.3M parameters, compared to ≈15M in [18]. Another key comparison is with KiU-Net, which learns a representation at 8× the original input spatial resolution. Unlike us, KiU-Net [27] uses max pooling layers, which discards information, to reduce the size of their high resolution representation to the original (1×) resolution. Whereas, we maintain the 2× resolution of our embedding until the output layer, which yields subpixel predictions that are aggregated by our learnable downsam-

Table 3. *Evaluation on small lesion subset.* While [31] achieves the highest precision, we note they have the second lowest recall out of all methods – missing small lesions can negatively impact patient recovery. In contrast, our method ranks second in precision and first across all other metrics.

Method	DSC	IOU	Precision	Recall
U-Net [19]	0.368	0.225	0.440	0.316
D-UNet [32]	0.265	0.180	0.377	0.264
CLCI-Net [31]	0.246	0.178	**0.662**	0.215
KiU-Net [27]	0.246	0.255	0.466	0.206
X-Net [18]	0.306	0.213	0.546	0.268
SPiN(Ours)	**0.424**	**0.269**	0.546	**0.347**

Table 4. *Ablation study on ATLAS.* Removing SPG and/or LD results in performance decrease (rows 1, 2, 6), and SPG cannot be substituted with more parameters or interpolation (rows 3–5). The best results are achieved by our full model (row 8).

Method	DSC	IOU	Precision	Recall
Without SPG, LD (Baseline)	0.634	0.487	0.707	0.606
Without SPG	0.637	0.487	0.701	0.613
Replace SPG with addit. convolutions	0.627	0.475	0.721	0.596
Replace SPG w/bilinear upsampling	0.663	0.513	0.780	0.600
Replace SPG w/nearest upsampling	0.660	0.513	0.762	0.626
Replace LD with downsampling	0.670	0.526	0.786	0.625
Full model with soft Dice loss	0.684	0.546	0.729	0.672
Full model	**0.703**	**0.556**	**0.806**	**0.654**

pler to the 1× resolution. Admittedly, this comes at the cost of runtime – our method requires 2.145 s per patient and KiU-Net [27] requires 1.05 s. However, we outperform [27] by an average of 33.7% across all metrics and reduce test time memory by half. We show qualitative comparisons in row 1 of Fig. 3 where the segmentation produced by our approach better captures irregularly shaped lesions than those predicted by competing methods.

Small Lesion Segmentation. Here, we consider the task of segmenting lesions that occupy fewer than 100 pixels or 0.2% of the image. Due to the challenging nature of the task, we observe an expected drop in performance across all methods (trained on the proposed split) when segmenting small lesions (Table 3), as compared to doing so for all lesion sizes (Table 2). However, we still outperform all competing methods – by even larger margins than on the full test set. This shows that competing methods, while able to localize large and medium sized lesions, actually perform poorly on small lesions. With the exception of preci-

sion, where we tie for second with X-Net [18], we rank first in all other metrics. We note that while CLCI-Net [31] has the highest precision, it also achieved second lowest recall, meaning that it misses many small lesions, which is critical to clinical prognosis and thus patient recovery. This is also reflected in DSC and IOU where we outperform [31] by 72% and 51%, respectively. Qualitatively, rows 2 and 3 in Fig. 3 show that our method successfully localized small lesions that [18, 19, 27, 31, 32] missed entirely.

Ablation Studies. Table 4 shows the effect of each of our contributions to architectural design. Row 1 shows that our baseline, a U-Net [19] based encoder-decoder, performs significantly worse by 11.7% than the proposed approach because it lacks fine local details from SPG and uses bilinear downsampling instead of a learnable downsampler (LD). Including LD alone, but not SPG (row 2) provides no improvement as the network only learns a coarse global representation, but is still missing details lost during spatial downsampling.

In row 3, we show that solely increasing parameters (i.e. adding ResNet blocks [7] to the baseline) brings no improvement, which suggests that the performance boost is *not* a result of a larger network. In fact, SPG and the learnable downsampler marginally increase the model size as they only combine for 27K parameters. Rows 4 and 5 show that using hand-crafted $2\times$ resolution images (from bilinear, nearest neighbor upsampling) does provide some gain. In these experiments, we replace SPG with different interpolation methods and the higher resolution images undergo 3×3 convolutions before being passed as skip connections to the decoder. However, because the $2\times$ representation is not learned, as it is with SPG, the result is still $\approx6\%$ worse than our full model. Our learnable downsampler (LD) contributes 4.4% to our performance (row 6) as removing LD and replacing it with bilinear interpolation smooths lesion boundaries, resulting in loss of details. Finally, we justify the use of cross entropy for our loss function; row 7 demonstrates that minimizing a soft Dice loss, as in [31, 32], results in worse performance. The best performance is achieved with our full model using SPG and LD, and minimizing cross entropy (row 8).

5   Discussion

We propose SPiN, a network architecture that learns a spatially increasing embedding that, when used as guidance for an encoder-decoder network, helps ensure that small structures are not lost through spatial downsampling in the encoder. We note that our embedding does not create extra spatial information (data processing inequality), but serves as a means for better characterization of local regions for the downstream segmentation task. While we outperform existing works and improve on small lesion segmentation, we do cost more memory and compute than the baseline. However, the extra cost is within reason (1 GB of memory for training and ≈ 0.7 s in runtime) and does not limit applicability. Despite the improved segmentation performance, we would like to address that there is still room for improvement, especially with small lesions. The highest

recall of 0.347 achieved by our model is admittedly low compared to recall metrics on the full dataset, implying that many small lesions still pass undetected. We note that this is one of the first works to study subpixel architectures in lesion segmentation, and we hope our optimistic results will motivate further exploration in this direction.

References

1. Asadi-Aghbolaghi, M., Azad, R., Fathy, M., Escalera, S.: Multi-level context gating of embedded collective knowledge for medical image segmentation. arXiv preprint arXiv:2003.05056 (2020)
2. Atlantis, T., et al.: Association of outcome with early stroke treatment: pooled analysis of atlantis, ecass, and ninds rt-pa stroke trials. Lancet **363**(9411), 768–774 (2004)
3. Chen, Y., Li, J., Xiao, H., Jin, X., Yan, S., Feng, J.: Dual path networks. arXiv preprint arXiv:1707.01629 (2017)
4. Ciresan, D., Giusti, A., Gambardella, L., Schmidhuber, J.: Deep neural networks segment neuronal membranes in electron microscopy images. Adv. Neural. Inf. Process. Syst. **25**, 2843–2851 (2012)
5. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 184–199. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_13
6. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2015)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
8. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
9. Hui, H., Zhang, X., Li, F., Mei, X., Guo, Y.: A partitioning-stacking prediction fusion network based on an improved attention u-net for stroke lesion segmentation. IEEE Access **8**, 47419–47432 (2020)
10. Kingma, D.P., Ba, J.L.: Adam: a method for stochastic gradient descent. In: ICLR: International Conference on Learning Representations, pp. 1–15 (2015)
11. Liew, S.L., et al.: A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. Sci. Data **5**, 180011 (2018)
12. Liew, S.L., et al.: The anatomical tracings of lesions after stroke (atlas) dataset-release 1.1. bioRxiv, p. 179614 (2017)
13. Maier, O., et al.: Isles 2015 - a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Med. Image Anal. **35**, 250–269 (2017)
14. Manvel, A., Vladimir, K., Alexander, T., Dmitry, U.: Radiologist-level stroke classification on non-contrast CT scans with deep U-net. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11766, pp. 820–828. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_91

15. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. Adv. Neural. Inf. Process. Syst. **32**, 8026–8037 (2019)
16. Pham, C.H., Ducournau, A., Fablet, R., Rousseau, F.: Brain MRI super-resolution using deep 3D convolutional networks. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pp. 197–200. IEEE (2017)
17. Pham, C.H., et al.: Multiscale brain MRI super-resolution using deep 3D convolutional networks. Comput. Med. Imaging Graph. **77**, 101647 (2019)
18. Qi, K., et al.: X-net: brain stroke lesion segmentation based on depthwise separable convolution and long-range dependencies. In: Shen, D., et al. (eds.) X-net: brain stroke lesion segmentation based on depthwise separable convolution and long-range dependencies. LNCS, vol. 11766, pp. 247–255. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_28
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
20. Rueda, A., Malpica, N., Romero, E.: Single-image super-resolution of brain MR images using overcomplete dictionaries. Med. Image Anal. **17**(1), 113–132 (2013)
21. Seyedhosseini, M., Sajjadi, M., Tasdizen, T.: Image segmentation with cascaded hierarchical models and logistic disjunctive normal networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), December 2013
22. Shi, W., et al.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1874–1883 (2016)
23. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.C.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems, vol. 28, pp. 802–810 (2015)
24. Song, H., Wang, W., Zhao, S., Shen, J., Lam, K.M.: Pyramid dilated deeper ConvLSTM for video salient object detection. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 715–731 (2018)
25. Tang, Z., Pan, B., Liu, E., Xu, X., Shi, T., Shi, Z.: SRDA-net: super-resolution domain adaptation networks for semantic segmentation. arXiv preprint arXiv:2005.06382 (2020)
26. Tureckova, A., Rodríguez-Sánchez, A.J.: ISLES challenge: U-shaped convolution neural network with dilated convolution for 3D stroke lesion segmentation. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11383, pp. 319–327. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11723-8_32
27. Valanarasu, J.M.J., Sindagi, V.A., Hacihaliloglu, I., Patel, V.M.: KiU-net: towards accurate segmentation of biomedical images using over-complete representations. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12264, pp. 363–373. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_36
28. Virani, S.S., et al.: Heart disease and stroke statistics-2020 update: a report from the American heart association. Circulation **141**(9), e139–e596 (2020)
29. Wang, L., Li, D., Zhu, Y., Tian, L., Shan, Y.: Dual super-resolution learning for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3774–3783 (2020)
30. Chan, B.P., Albers, G.W.: Acute ischemic stroke. Curr. Treat. Options. Neurol. **1**(2), 83–95 (1999). https://doi.org/10.1007/s11940-999-0009-5

31. Yang, H., et al.: CLCI-net: cross-level fusion and context inference networks for lesion segmentation of chronic stroke. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11766, pp. 266–274. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_30
32. Zhou, Y., Huang, W., Dong, P., Xia, Y., Wang, S.: D-Unet: a dimension-fusion u shape network for chronic stroke lesion segmentation. IEEE/ACM Trans. Comput. Biol. Bioinform. (2019)