Weakly-Supervised Convolutional Neural Networks for Vessel Segmentation in Cerebral Angiography

Arvind Vepa[1], Andrew Choi[1], Noor Nakhaei[1], Wonjun Lee[1], Noah Stier[2], Andrew Vu[3], Greyson Jenkins[4], Xiaoyan Yang[1], Manjot Shergill[1], Moira Desphy[1], Kevin Delao[1], Mia Levy[1], Cristopher Garduno[1], Lacy Nelson[1], Wandi Liu[1], Fan Hung[1], and Fabien Scalzo[1,4]

[1]University of California, Los Angeles, [2]University of California, Santa Barbara, [3]Cal State University, Fullerton, [4]Pepperdine University

Abstract

Automated vessel segmentation in cerebral digital subtraction angiography (DSA) has significant clinical utility in the management of cerebrovascular diseases. Although deep learning has become the foundation for state-of-the-art image segmentation, a significant amount of labeled data is needed for training. Furthermore, due to domain differences, pretrained networks cannot be applied to DSA data out-of-the-box. To address this, we propose a novel learning framework, which utilizes an active contour model for weak supervision and low-cost human-in-the-loop strategies to improve weak label quality. Our study produces several significant results, including state-of-the-art results for cerebral DSA vessel segmentation, which exceed human annotator quality, and an analysis of annotation cost and model performance trade-offs when utilizing weak supervision strategies. For comparison purposes, we also demonstrate our approach on the Digital Retinal Images for Vessel Extraction (DRIVE) dataset. Additionally, we will be publicly releasing code to reproduce our methodology and our dataset, the largest known high-quality annotated cerebral DSA vessel segmentation dataset.

1. Introduction

Digital subtraction angiography (DSA) is the gold standard in vessel visualization and extremely important in the diagnosis and treatment of arterial and venous occlusions. Vessel segmentation in cerebral DSA is crucial to the management of cerebrovascular diseases such as stroke diagnosis and detection of aneurysms. While manual annotation is possible, it is often elaborate and complex, limiting the study of large cohorts of images [10]. Consequently, automated vessel segmentation methods have become increasingly necessary.

With recent advances in deep learning, convolutional neural networks (CNN) have shown substantial improvement in medical image segmentation. To the best of our knowledge, our work is the first to incorporate active contour models as weak supervision for semantic segmentation deep learning training. By doing so, we significantly reduce the human cost related to annotation generation. The rest of the paper is as follows. In Section 2, we review prior work in vessel segmentation, weak supervision, and deformable models. In Section 3, we propose our methodology for weak supervision. Finally, in Section 4, we employ our methodology on the datasets and discuss the results and implications.

2. Related Work

Several methods have been proposed to segment vessels in DSA; however, they all suffer from limitations associated with noise, bone artifacts, significant vessel diameter differences, and small dataset sizes [10, 5, 25, 37]. Despite limited research in DSA vessel segmentation, there has been significant research in blood vessel segmentation in fundus retinal images largely due to the availability of several public data sets. Recent studies have used deep learning methods such as multi-scale and multi-level CNNs [9], multi-path supervision [31], deformable CNNs [15], and feature pyramid cascade networks [30].

Due to the the difficulty of acquiring labeled data, there has recently been considerable interest in weak supervision for CNNs for medical image segmentation. Prior work have incorporated weak supervision into CNNs for histopathology image segmentation[14] as well as brain tumor segmentation [13] with methods for weak supervision including bounding boxes, scribbles, image level tags, and partial labels [4, 32]. Human-in-the-loop strategies have also been

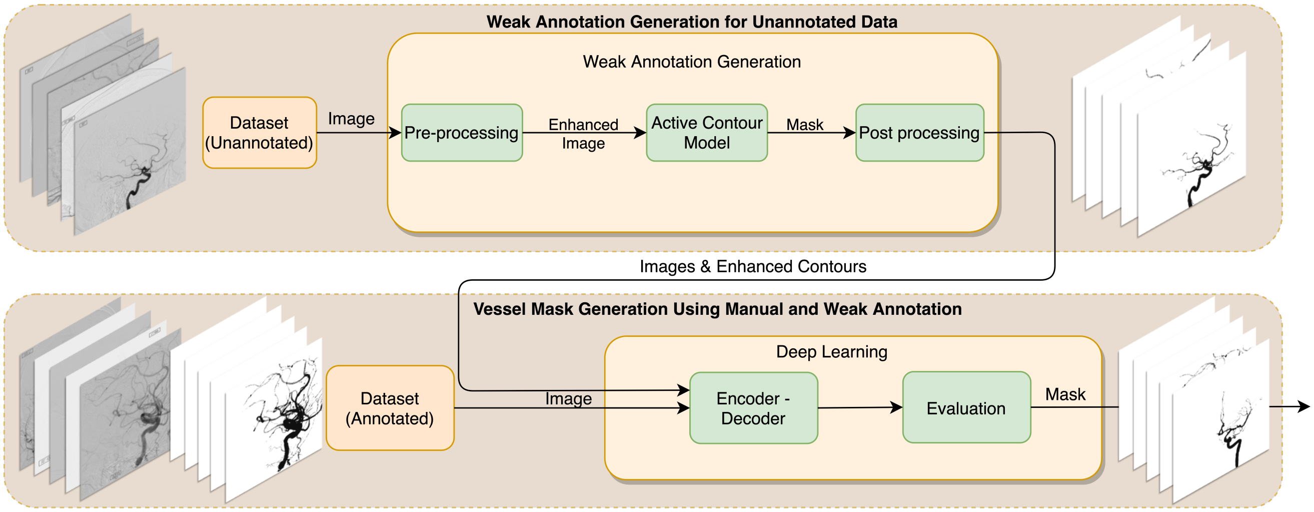

Figure 1: Pipeline of our approach which combines manual and weak annotations.

explored for active learning [34, 2].

Generalizing these weak supervision approaches to blood vessels is nontrivial as blood vessels often possess a significantly more complex and intricate structure (see Figure 3) than the objects of focus in prior work. In one study, hierarchical clustering was used as weak annotations and weak annotations from different but related domains were used to augment the training set for retinal segmentation [20]. In another study, vessel enhancement algorithms were used to generate weak annotations and an iterative optimization process and active learning were used to improve annotation quality [36]. Methods such as this use significant annotation and optimization resources. These considerations make utilizing relatively computationally cheap deformable models an attractive choice.

Deformable models such as snakes and level-set method are pioneering methods in autonomous segmentation of 2D and 3D medical images [22, 24, 35]. Recently, several researchers have proposed methods combining the strengths of deformable models and CNNs in the automated segmentation of left ventricles [3] and cardiac walls [33]. Although these works combine both deformable models and deep learning in their approach, the deformable models are primarily used as refinement of deep learning segmentation predictions rather than weak supervision for deep learning training.

With this in mind, our work is the first to use active contour models as weak supervision for CNN semantic segmentation model training. By utilizing large unlabeled datasets through computationally efficient active contour methods with limited human intervention, our proposed method shows extremely promising results for CNN segmentation despite the intricacy of blood vessel structures.

3. Methodology

In our proposed approach, we obtain several different types of annotations for supervised training of a semantic segmentation deep learning model. We use active contour models to generate weak annotations with limited human intervention and compare our approach to popular baseline methods for automated segmentation. Additionally, we combine weak annotations with manual annotations in order to assess whether we can improve model performance with lower annotation costs. For each set of annotations, we train a semantic segmentation model and evaluate its performance on a held out test set. Figure 1 displays a pipeline of our approach which combines manual and weak annotations. Similar to prior work [36], our approach is compatible with any CNN-based segmentation architecture.

3.1. Datasets

The Cerebral DSA imaging dataset utilized in this study was obtained from patients evaluated at a comprehensive stroke center and identified with symptoms of acute ischemic stroke. Inclusion criteria for this study included: (1) final diagnosis of acute ischemic stroke, (2) last known well time within six hours at admission, and (3) DSA of the brain performed as part of a thrombectomy procedure. A total of 91 patients satisfied the above criteria and were included in this study. The DSA scanning was performed on a Philips Allura Xper FD20® Biplane using a routine timed contrast-bolus passage technique. A manual injection of omnipaque 300 was performed at a dilution of 70% (30% saline) such that 10cc of contrast was administered intravenously at an approximate rate of 5cm^3/s with median peak voltage output of 95 Kv, IQR 86, 104. Image sizes were all 1024×1024 but were acquired with different fields of view [10]. The total dataset consists of 128 images which was partitioned into a $75-25$ training test split.

In order to compare our approach to other published results, we also replicate our results on DRIVE. This dataset consists of a total of 40 color fundus photographs acquired from a diabetic retinopathy screening program. Images were acquired using a Canon CR5 non-mydriatic 3CCD camera with a 45° field of view (FOV). Each image is captured using 8 bits per color plane at 768 × 548 pixels. The FOV of each image is circular with a diameter of approximately 540 pixels. The set of 40 images is divided into a test and training set both containing 20 images [26].

3.2. Manual Annotations (MA)

To obtain the manual annotations for the Cerebral DSA dataset, fifteen observers were trained by research staff which included experienced stroke researchers. All manual annotations were developed using GIMP, a popular open-source raster graphics editor. Twelve observers annotated the training set with no overlap while two observers annotated the test set. The Cohen's kappa coefficient - a statistic to measure inter-rater reliability - was 0.799 which demonstrates that the annotation quality was highly reliable. The final observer annotated a subset of the test set to determine which set of test annotations would be the gold standard ground truth based on inter-annotator agreement. The remaining set of annotations (referred to as ”second annotator” annotations) was used as an additional datapoint in our performance comparison. To standardize for differing skill levels, annotation times were only collected from those who generated both manual and weak annotations. With this, the mean annotation time per image was 1.92 hours. The DRIVE dataset manual annotations are publicly available. Three observers were trained by an ophthalmologist and manually segmented the dataset. The first observer and second observer annotated both the train and test set. A third observer separately annotated the test to compare the performance of an independent human observer with automated methods. The Cohen's kappa coefficient was 0.759 which demonstrates that the annotation quality was highly reliable [26].

3.3. Active Contour Model

For our active contour model, we use a hybrid level set model (CGLI) proposed by Chen et al. [7] based on the Selective Binary and Gaussian Filtering Regularized Level Set (SBGFRLS) model and the local fitting term in Local Binary fitting (LBF) model for retina vessel segmentation. We chose this model for three primary reasons: (1) the signed pressure force function introduced by the SBGFRLS model combined with the local intensity property introduced by the LBF model allows for proper segmentation of low contrast vessels, (2) the approach is more robust to initial conditions than traditional methods, and (3) strong results have been demonstrated on two other vessel imaging datasets, DRIVE and STARE.

Let $\Omega \subset \mathbb{R}^2$ be the image space and $I : \Omega \rightarrow \mathbb{R}$ be an intensity of the image space. The CGLI model consists of a local force functional g^L that captures the intensity difference between the inside and outside of the contours and a global force functional g^G that is used to avoid the model being trapped in a local minima. More specifically, the local force is defined as

$$g^L(I(x)) = \frac{\int K_\sigma(y-x)(I(x)-A(y))dy}{\max(|\int K_\sigma(y-x)A(y)dy|)} \tag{1}$$

where K_σ is a smoothing kernel function with variance σ, and $A(y) = 0.5(f_1(y) + f_2(y))$.

Here, $f_1(x)$ and $f_2(x)$ are two spatial dependent functions that approximate the intensities of the inside and outside of the contours near the point $x \in \Omega$ and enable the algorithm to detect the accurate segmentation in the intensity inhomogeneity more efficiently than the vanilla level set method. They are defined as

$$\begin{aligned} f_1(x) &= \frac{\int K_\sigma(y-x)I(x)H_\epsilon(\phi(y))dy}{\int K_\sigma(y-x)H_\epsilon(\phi(y))dy} \\ f_2(x) &= \frac{\int K_\sigma(y-x)I(x)(1-H_\epsilon(\phi(y)))dy}{\int K_\sigma(y-x)(1-H_\epsilon(\phi(y)))dy} \end{aligned} \tag{2}$$

where ϕ is the signed distance function and $H_\epsilon = 0.5 + \arctan(x/\pi)/\pi$ is a smooth Heaviside step function.

The global force is then defined as

$$g^G(I(x)) = \frac{I(x)-(c_1+c_2)/2}{\max(|I(x)-(c_1+c_2)/2|)} \tag{3}$$

where c_1 and c_2 are global average intensity values.

Finally, using both the local and global force, the hybrid level set method computes the signed distance function that satisfies the following partial differential equation:

$$\frac{\partial\phi(t,x)}{\partial t} = (g^L(I(x)) + \omega g^G(I(x))) \cdot \alpha|\nabla\phi(t,x)| \tag{4}$$

where α controls the evolution speed and ω regulates the influence between the local and global forces. The algorithm of the hybrid CGLI level set method is shown in Alg. 1.

Algorithm 1 Hybrid CGLI Level Set Method

```
Input: An image intensity function $I$
Output: A signed distance function $\phi$
Initialize $\phi \leftarrow K_\sigma * I - \frac{\max(I)+\min(I)}{2}$
for $i = 1, \dots, N$ do
    for $x \in \Omega$ do
        if $\phi(x) > 0$ then
            $\phi(x) \leftarrow 1$
        else
            $\phi(x) \leftarrow -1$
    $\phi \leftarrow \phi + \Delta_t(g^L + \omega g^G)\sqrt{(\partial_x \phi)^2 + (\partial_y \phi)^2}$
```

In our DSA experiments we use the heat equation and in our DRIVE experiments we use a Gaussian kernel for the smoothing kernels. Furthermore, we use the forward difference scheme to solve for derivatives $\partial_x \phi$ and $\partial_y \phi$.

3.3.1 Fully-automated Active Contour Weak Annotations (FACWA)

To generate FACWA, the DSA image is pre-processed using a Gaussian filter with $\sigma_p = 1.0$ and passed to the active contour model which runs for 600 iterations with optimal parameters of $\sigma = 1\text{e}{-7}$, $\omega = 0.2$, and $\epsilon = 0.2$. In order to faithfully represent the lack of manual annotation data, optimal hyper-parameters were chosen based on visual comparison. The following hyper-parameters were considered: $\sigma \in [1\text{e}{-6}, 1\text{e}{-7}]$, $\omega \in [0.1, 0.2, 0.3, 0.5, 0.7]$, and $\epsilon \in [0.01, 0.1, 0.2, 0.3, 0.5]$. The model consumes 260 MB of RAM and runs for approximately 200 seconds per image. We looked at slightly different parameters for the DRIVE dataset and tuned the parameters differently for better comparison with other related work on the DRIVE dataset.

3.3.2 Human-in-the-Loop Active Contour Weak Annotations (HACWA)

A universal pipeline for automated segmentation is often inadequate as images within the same dataset can differ drastically in terms of quality and clarity. For the DSA dataset, we address these issues by generating several annotation candidates by changing pre-processing parameters. We then use human intervention to choose the best segmentation based on a visual comparison of the generated annotations. For pre-processing, we use a combination of Gaussian blurring followed by contrast limited adaptive histogram equalization (CLAHE). We vary σ_p in Gaussian blur as well as the clip limit δ in CLAHE while the tile size of CLAHE is set constant as 8×8. We use $\sigma_p \in [1.0, 1.5, 2.0]$ for low blurring, medium blurring, and high blurring and $\delta \in [2.0, 4.0]$ for low contrast and high contrast. The combination of these parameters produce six total annotation candidates for each image. The user then picks the best one or discards all of them if none are of sufficient quality. Using this method, a total of 50 images from the training set were deemed to have annotations of sufficient quality. The mean annotation time per image was 1.02 minutes which is drastically lower than the annotation time for MA.

3.3.3 Baseline Weak Annotations

For the baseline fully-automated weak annotations (BFWA), we chose a popular light-weight unsupervised image segmentation method, kernel graph-cut segmentation [27]. The hyper-parameters considered were $\alpha \in [0.0, 0.01, 0.1, 0.25, 0.5]$, $a \in [1, 2, 3]$, $b \in [0.5, 1.0, 1.5]$, and $\sigma_b \in [0.25, 0.5, 0.75]$ and the optimal values were chosen as $\alpha = 0.0$, $a = 2$, $b = 1.0$, and $\sigma_b = 0.5$, based on visual comparison.

For the baseline human-in-the-loop weak annotations (BHWA), we used the popular fuzzy select tool implemented by GIMP to segment vessels. To use the tool, the user selects a seed point within the image, which results in adjacent pixels with a similar intensity to the selected pixels being selected. Users can modulate a similarity threshold for the intensity of the selected pixels [1]. The mean annotation time per image for this method was 1.08 minutes.

3.4. Supervised Model Training

In order to compare the different annotation approaches, we train and evaluate a deep learning semantic segmentation model that utilizes the annotations for supervised learning. The experiments were conducted on Amazon EC2 instances and utilized the g4dn.xlarge instances. The CNNs were implemented in PyTorch 1.6.0. A total of 4.9 GB and 1.7 GB of GPU memory and 4.3 GB and 3.0 GB of RAM are consumed during model training and inference with training and inference time of 0.38 and 0.44 seconds per image respectively.

3.4.1 Encoder-Decoder Model Architecture

In order to stress the generality of our approach, we utilized popular and publicly available encoder-decoder architectures for our segmentation model. The encoder networks were pre-trained on ImageNet. We tested several notable encoder [11, 28, 12, 29] and decoder [18, 8, 17, 38, 6] networks and determined that the DenseNet169 encoder and Feature Pyramid Network decoder were optimal for the DSA dataset and VGG16 encoder and UNet decoder were optimal for the DRIVE dataset.

3.4.2 Deep Learning Model Training

To optimize the loss function, we trained for 150 epochs for the DSA dataset and 1000 epochs for DRIVE using the Adam algorithm, a learning rate of 0.0001, and 0.20 dropout

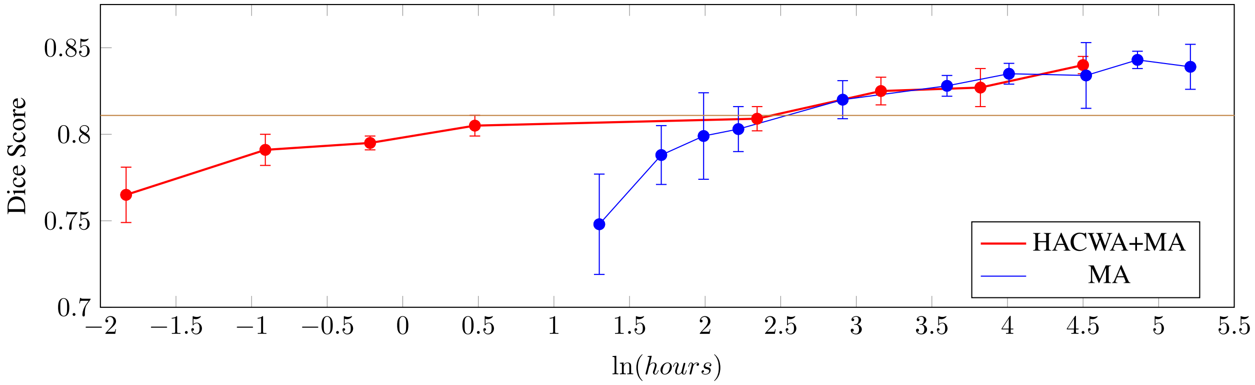

Figure 2: The graph describes the relationship between model performance and annotation time for models trained with HACWA+MA (red line) and models trained exclusively with MA (blue line). Each filled circle represents a datapoint. The brown horizontal line represents the second annotator score on the test set. For the first four datapoints on the red line, the models are trained exclusively with HACWA. For subsequent datapoints on the red line, the models are trained with 100% HACWA and proportions of MA.

applied to the last layer. In order to correct for class imbalance for the DSA dataset (this was not applied to DRIVE), we used weighted binary cross entropy (WBCE) as our loss. For the WBCE positive class weight, we tested values of 1.0 and 0.50 times the ratio of negative and positive samples (optimal was 1.0). During training we applied several image augmentations, including horizontal flip, shift, and Gaussian noise.

3.4.3 Hyperparameter Testing

For hyperparameter tuning, we first optimized the hyperparameters on a model trained exclusively on manual annotations. The hyperparameters for all other experiments were then fixed (except for the number of epochs which we increased to 300 for small samples for the DSA dataset). This was done to demonstrate parameter robustness for the different annotation types. To obtain the model hyperparameters, we created three non-overlapping validation sets from the training set to generate three training and validation splits. We then used grid search to generate hyperparameter combinations and trained and validated each combination on each split. The performances were then averaged on each fold to determine optimal parameters.

3.4.4 Evaluation Metrics

To evaluate the performance of our model on the DSA dataset, we used three metrics: DICE, Average Precision (AP), and Area Under the Curve of the Receiver Operating Characteristic (AUROC). DICE is a commonly used evaluation metric and is defined as:

$$DICE = \frac{2TP}{2TP + FP + FN} \tag{5}$$

Unlike Accuracy, DICE considers low prevalence of the positive class, which is why it is a popular metric for segmentation. AP is also a commonly used metric and is defined as the area under the Precision-Recall curve. AUROC is calculated by obtaining the area under the Receiver Operating Characteristic, which is generated by plotting the True Positive rate against the False Positive rate. Both AP and AUROC consider the model performance on several decision thresholds. AP is generally preferred to AUROC in cases with class imbalance though we report AUROC because AUROC is often more commonly reported. For the DRIVE dataset, we report Accuracy and AUROC for comparison purposes because they are the most commonly reported for this dataset. Unless otherwise stated, metrics are only generated from the FOV for DRIVE. Thresholded metrics, like DICE and Accuracy, are more relevant for clinical use.

4. Experiments

We trained supervised models on the different annotations and considered performance of models trained on varying amounts and combinations of annotations. For models trained on a mix of weak annotations and MA, for the DSA dataset we trained a supervised model on a combination of 100% of HACWA and varying amounts of MA for the images that were excluded from HACWA. For the DRIVE dataset, a certain proportion of weak annotations were replaced with MA. For each experiment on the DSA dataset, we trained the model ten times with ten different random seeds, which affected the training data and parameter initialization, and reported statistics for test set metrics over all the models. For the DRIVE dataset, because of the high-quality FACWA, we did not generate HACWA.

Approach	DICE (%)	Average Precision (%)	AUROC (%)	Annotation Time (hours)	Optimization Time (hours)
Combined Weakly-supervised and Supervised					
100% HACWA+48% MA	**84.0 ± 0.5**	92.2 ± 0.4	98.9 ± 0.2	89.95	1.5
100% HACWA+24% MA	**82.7 ± 1.1**	91.6 ± 0.5	98.0 ± 0.1	45.86	1.2
100% HACWA+11% MA	**82.5 ± 0.8**	90.8 ± 1.1	98.4 ± 0.1	21.90	0.97
100% HACWA+4% MA	80.9 ± 0.7	89.8 ± 0.5	98.4 ± 0.2	9.00	0.87
Weakly-supervised					
100% HACWA	80.5 ± 0.6	89.5 ± 0.6	98.3 ± 0.3	1.6	0.80
50% HACWA	79.5 ± 0.4	88.9 ± 0.7	97.9 ± 0.6	0.81	0.40
25% HACWA	79.1 ± 0.9	88.4 ± 0.8	98.0 ± 0.3	0.41	0.20
100% FACWA	76.3 ± 0.8	84.9 ± 1.3	96.1 ± 1.2	0	1.5
Supervised					
100% MA	**83.9 ± 1.3**	92.8 ± 0.4	99.0 ± 0.1	184.3	1.5
70% MA	**84.3 ± 0.5**	92.3 ± 0.4	98.8 ± 0.2	129.0	1.1
50% MA	**83.4 ± 1.9**	92.2 ± 0.5	98.9 ± 0.2	92.2	0.77
30% MA	**83.5 ± 0.6**	91.9 ± 0.4	98.7 ± 0.3	55.3	0.46
20% MA	**82.8 ± 0.6**	91.6 ± 0.4	98.8 ± 0.1	36.9	0.31
10% MA	82.0 ± 1.1	90.9 ± 0.5	98.6 ± 0.3	18.4	0.31
5% MA	80.3 ± 1.3	89.9 ± 0.6	98.7 ± 0.2	9.22	0.15
4% MA	79.9 ± 2.5	89.5 ± 1.0	98.7 ± 0.1	7.37	0.12
3% MA	78.8 ± 1.7	87.5 ± 2.1	98.4 ± 0.3	5.53	0.093
2% MA	74.8 ± 2.9	84.2 ± 3.6	97.8 ± 1.0	3.69	0.062
Baseline					
100% BHWA	80.0 ± 1.0	87.5 ± 1.0	98.1 ± 0.3	1.7	1.5
50% BHWA	79.3 ± 1.1	87.6 ± 1.0	97.8 ± 0.5	0.87	0.77
25% BHWA	78.0 ± 1.0	85.9 ± 1.8	97.9 ± 0.4	0.43	0.39
100% BFWA	60.3 ± 1.9	69.0 ± 2.7	91.3 ± 2.7	0	1.5
Annotations					
Second Annotator*	81.1	X	X	X	X
FACWA*	79.0	X	X	X	X
BFWA*	59.3	X	X	X	X

Table 1: These are the metric scores for the models trained on the different annotations on the DSA dataset. The rows indicated by (*) are the scores for the annotations on the test set. Using a significance level of 5% and the Bonferroni correction for multiple comparisons, the DICE scores which are significantly greater than the second annotator are bolded.

4.1. Results

The metric results from our experiments on the DSA dataset can be seen in Table 1. Our models trained on FACWA performed better than BFWA. However, all the models trained on HACWA perform better than the model trained on the FACWA, which demonstrates that limited human supervision has a significant impact on model performance. Additionally, models trained on HACWA produced much higher performance than models trained on MA given the annotation time. For example, the models trained on 100% HACWA obtained near human-level performance given an annotation time of 1.6 hours for the training set. In contrast, models trained on 4% MA and 5% MA have similar performance to the models trained on 100% HACWA yet utilize 7.37 hours and 9.22 hours, corresponding to five and six times the annotation time respectively. Additionally, the models trained on 4% MA and 5% MA have higher variance in model results because of the small dataset sizes, which suggests more variability in annotation quality.

In Figure 3, results from models trained on 100% FACWA, 100% BFWA, 100% HACWA, 100% BHWA, 100% HACWA+ 48% MA, and 100% MA on two test images are shown. The second image has significantly more noise than the first image yet all the models are able to successfully discriminate between the noise and blood vessels. The quality of the predictions from the model trained on 100% FACWA are significantly better than the model trained on 100% BFWA and are able to capture much more of the vessel structure. The quality of the predictions from

Approach	Accuracy (%)	AUROC (%)
Supervised		
Jin *et al*. [15]	95.7	98.0
Liskowski and Krawiec [19]	95.4	97.9
Maji *et al*. [21]	94.7	92.8
Marin *et al*. [23]	94.5	95.9
Weakly- and Unsupervised		
Lu *et al*. ° [20]	95.6	95.8
Neimeijer *et al*. [26]	93.8	89.8
Kande *et al*. [16]	89.1	95.2
Our Approach		
100% MA	95.6	96.5
75% FACWA+25% MA	94.7	93.0
100% FACWA	94.2	92.1
100% FACWA°	95.7	94.2
Annotations		
Second Annotator*	94.7	X
FACWA*	92.4	X

Table 2: Metric scores for previous work and our models trained on different annotations on the DRIVE dataset. The rows indicated by (*) are the scores for the annotations on the test set. The rows indicated by (°) are the scores generated from the entire image rather than just the FOV.

the model trained on 100% HACWA are slightly better than the model trained on 100% FACWA and 100% BHWA. Visually, predictions from models trained on HACWA+ 48% MA and 100% MA are very similar from the model trained on 100% HACWA. All three of the models are effectively able to discriminate between the blood vessels and noise and capture the vast majority of the intricate blood vessel structure.

In Figure 2, we show the relationship between model performance and annotation time for models trained with HACWA+MA and models trained exclusively with MA. For the first four datapoints on the red line, the models are trained exclusively with HACWA and perform far better for the given annotation cost than the models trained exclusively on MA. For subsequent datapoints on the red line, the models are trained with 100% HACWA and proportions of MA and are able to achieve human annotation quality with less annotation time than models trained exclusively on MA. Models trained on HACWA notably have very small error bars unlike models trained on MA which are susceptible to high variance with small sample sizes. Using a significance level of 5% and the Bonferroni correction for multiple comparisons, our analysis of the respective p-values strongly suggest HACWA+11% MA would result in model performance exceeding a human annotator while involving minimal additional annotation time. As can be seen in Table 1, models trained on various annotation combinations have statistically significantly greater performance than a human annotator, which has significant annotation cost implications. Model performance that surpasses inter-annotator agreement suggests that such models can generate annotations as reliably as humans, which enables a wide range of practical applications.

4.2. Comparison with Related Work

The metric results for previous work and our models trained on different annotations on the DRIVE dataset can be seen in Table 2. In terms of accuracy, our model trained on FACWA performed better than all the weakly- and unsupervised approaches. While our AUROC is lower than some models, thresholded predictions are more relevant for clinical use and strong AUROC which translates to poor accuracy is not clinically relevant (*e.g*. Kande *et al*. [16]). One of the competitive scores, Lu *et al*. [20], used hierarchical clustering as weak annotations. However, images from other retinal segmentation datasets were used to augment the training set whereas we were able to generate our results from the DRIVE training set images alone.

Additionally, the 94.2% accuracy obtained was very close to the human annotator accuracy of 94.7%. In fact, by only exchanging five FACWA with MA, the model is able to obtain human annotator performance. The accuracy obtained surpasses a few deep learning supervised methods (Marin *et al*. [23] and Maji *et al*. [21]), which is impressive because the model only uses a quarter of the annotation resources. Our supervised model has metric scores close to the most competitive supervised models, though enhancing our supervised model is out of the scope of the current study.

While weakly supervised learning in vessel segmentation is one of the focuses of this study, our study also produces pioneering work in DSA vessel segmentation. In comparison to other work, our dataset of 128 high-quality annotated DSA images is the largest used by far (two studies in this area used 30 annotated images [37, 25] and another one used 88 annotated images[10]). Based on a visual comparison of other DSA study results [37, 10], our models produce higher quality vessel masks and are able to better discriminate noise. A previous study which utilized data in the current study achieved an AUCROC of 95.1% [10], which is well below the model scores in our current study. We will be the first group to make our annotated DSA dataset publicly available. Not only will our results be the first that are completely reproducible, but the publicly available data will spur other researchers to continue work and improve methodologies in this area.

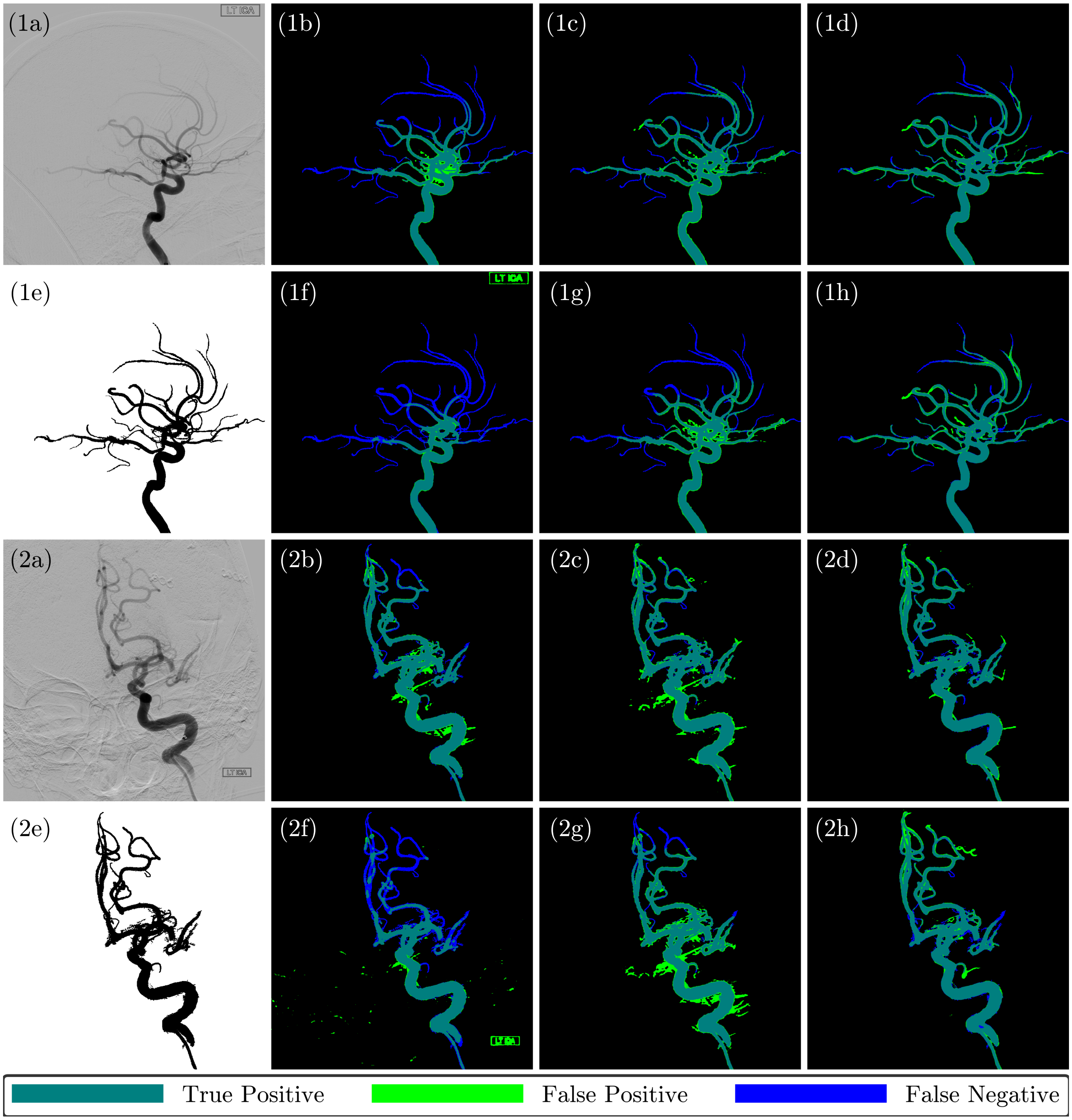

Figure 3: Two test images are shown with their respective (a) DSA image, (b) 100% FACWA, (c) 100% HACWA, (d) 100% HACWA+48% MA, (e) Ground Truth, (f) 100% BFWA, (g) 100% BHWA, and (h) 100% MA.

5. Conclusion

DSA vessel segmentation is a challenging problem; however, based on the results in our study, we were able to demonstrate significant practical advantages to using our weak learning framework in comparison to other related work. One weakness in our study is the generalization of our approach to additional datasets as well as different network architectures and computer vision tasks. In future work, we hope to address these issues.

References

[1] 2.5. fuzzy selection (magic wand).

[2] David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Efficient interactive annotation of segmentation datasets with polygon-rnn++. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 859–868, 2018.

[3] M. R. Avendi, Arash Kheradvar, and Hamid Jafarkhani. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. *CoRR*, abs/1512.07951, 2015.

[4] Junjie Bai and Xiaodong Wu. Error-tolerant scribbles based interactive image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 392–399, 2014.

[5] Marco Boegel, Philip Hoelter, Thomas Redel, Andreas Maier, Joachim Hornegger, and Arnd Doerfler. A fully-automatic locally adaptive thresholding algorithm for blood vessel segmentation in 3d digital subtraction angiography. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2006–2009. IEEE, 2015.

[6] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2017.

[7] Guannan Chen, Meizhu Chen, Jichun Li, and Encai Zhang. Retina image vessel segmentation using a hybrid cgli level set method. *BioMed research international*, 2017, 2017.

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[9] Huazhu Fu, Yanwu Xu, Stephen Lin, Damon Wing Kee Wong, and Jiang Liu. Deepvessel: Retinal vessel segmentation via deep learning and conditional random field. In *International conference on medical image computing and computer-assisted intervention*, pages 132–139. Springer, 2016.

[10] Yang Fu, Jiawen Fang, Benjamin Quachtran, Natia Chachkhiani, and Fabien Scalzo. Vessel detection on cerebral angiograms using convolutional neural networks. In *International Symposium on Visual Computing*, pages 659–668. Springer, 2016.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[13] Zhanghexuan Ji, Yan Shen, Chunwei Ma, and Mingchen Gao. Scribble-based hierarchical weakly supervised learning for brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 175–183. Springer, 2019.

[14] Zhipeng Jia, Xingyi Huang, I Eric, Chao Chang, and Yan Xu. Constrained deep weak supervision for histopathology image segmentation. *IEEE transactions on medical imaging*, 36(11):2376–2388, 2017.

[15] Qiangguo Jin, Zhaopeng Meng, Tuan D Pham, Qi Chen, Leyi Wei, and Ran Su. Dunet: A deformable network for retinal vessel segmentation. *Knowledge-Based Systems*, 178:149–162, 2019.

[16] Giri Babu Kande, P Venkata Subbaiah, and T Satya Savithri. Unsupervised fuzzy based vessel segmentation in pathological digital fundus images. *Journal of medical systems*, 34(5):849–858, 2010.

[17] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018.

[18] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[19] Paweł Liskowski and Krzysztof Krawiec. Segmenting retinal blood vessels with deep neural networks. *IEEE transactions on medical imaging*, 35(11):2369–2380, 2016.

[20] Zheng Lu, Dali Chen, and Dingyu Xue. Weakly supervised retinal vessel segmentation algorithm without groundtruth. *Electronics Letters*, 56(23):1235–1237, 2020.

[21] Debapriya Maji, Anirban Santara, Pabitra Mitra, and Debdoot Sheet. Ensemble of deep convolutional neural networks for learning to detect retinal vessels in fundus images. *arXiv preprint arXiv:1603.04833*, 2016.

[22] Ravi Malladi and James A Sethian. Level set methods for curvature flow, image enchancement, and shape recovery in medical images. In *Visualization and mathematics*, pages 329–345. Springer, 1997.

[23] Diego Marín, Arturo Aquino, Manuel Emilio Gegúndez-Arias, and José Manuel Bravo. A new supervised method for blood vessel segmentation in retinal images by using gray-level and moment invariants-based features. *IEEE Transactions on medical imaging*, 30(1):146–158, 2010.

[24] Tim McInerney and Demetri Terzopoulos. Medical image segmentation using topologically adaptable snakes. In *International Conference on Computer Vision, Virtual Reality, and Robotics in Medicine*, pages 92–101. Springer, 1995.

[25] Cai Meng, Kai Sun, Shaoya Guan, Qi Wang, Rui Zong, and Lei Liu. Multiscale dense convolutional neural network for dsa cerebrovascular segmentation. *Neurocomputing*, 373:123–134, 2020.

[26] Meindert Niemeijer, Joes Staal, Bram van Ginneken, Marco Loog, and Michael D Abramoff. Comparative study of retinal vessel segmentation methods on a new publicly available database. In *Medical imaging 2004: image processing*, volume 5370, pages 648–656. International Society for Optics and Photonics, 2004.

[27] Mohamed Ben Salah, Amar Mitiche, and Ismail Ben Ayed. Multiregion image segmentation by parametric kernel graph cuts. *IEEE Transactions on Image Processing*, 20(2):545–557, 2010.

[28] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*, 2016.

[29] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

[30] Wei Wang, Jiafu Zhong, Huisi Wu, Zhenkun Wen, and Jing Qin. Rvseg-net: An efficient feature pyramid cascade

network for retinal vessel segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 796–805. Springer, 2020.

[31] Yicheng Wu, Yong Xia, Yang Song, Donghao Zhang, Dongnan Liu, Chaoyi Zhang, and Weidong Cai. Vessel-net: retinal vessel segmentation under multi-path supervision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 264–272. Springer, 2019.

[32] Jia Xu, Alexander G Schwing, and Raquel Urtasun. Learning to segment under various forms of weak supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3781–3790, 2015.

[33] D. Yang, Q. Huang, L. Axel, and D. Metaxas. Multi-component deformable models coupled with 2d-3d u-net for automated probabilistic segmentation of cardiac walls and blood. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 479–483, 2018.

[34] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 399–407. Springer, 2017.

[35] Alistair A Young, Dara L Kraitchman, Lawrence Dougherty, and Leon Axel. Tracking and finite element analysis of stripe deformation in magnetic resonance tagging. *IEEE Transactions on Medical Imaging*, 14(3):413–421, 1995.

[36] Jingyang Zhang, Guotai Wang, Hongzhi Xie, Shuyang Zhang, Ning Huang, Shaoting Zhang, and Lixu Gu. Weakly supervised vessel segmentation in x-ray angiograms by self-paced learning from noisy labels with suggestive annotation. *Neurocomputing*, 417:114–127, 2020.

[37] Min Zhang, Chen Zhang, Xian Wu, Xinhua Cao, Geoffrey S Young, Huai Chen, and Xiaoyin Xu. A neural network approach to segment brain blood vessels in digital subtraction angiography. *Computer methods and programs in biomedicine*, 185:105159, 2020.

[38] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.