Semi-supervised Medical Image Segmentation via Feature-perturbed Consistency

Yang Yang[1] Ruixuan Wang[2,3] Tong Zhang[2] Jingyong Su[1*]

[1] School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), China
[2] Department of Network Intelligence, Peng Cheng Laboratory, Shenzhen, China
[3] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China

Abstract—**Although deep convolutional neural networks have achieved satisfactory performance in many medical image segmentation tasks, a considerable annotation challenge still needs to be solved, which is expensive and time-consuming for radiologists. Most existing popular semi-supervised methods mainly impose data-level perturbations (e.g., rotation, noising) or feature-level perturbations (e.g., MC dropout) on unlabeled data. In this paper, we propose a novel semi-supervised segmentation strategy with meaningful perturbations at the feature level to leverage abundant useful information naturally embedded in the unlabeled data. Specifically, we develop a dual-task network where the segmentation head produces multiple predictions with a perturbation module, and the reconstruction head further utilizes the semantic information to enhance segmentation performance. The proposed framework subtly perturbs the network at the feature-level to generate predictions which should be similar and consistent. However, enforcing them roughly to be consistent at all pixels harms stable training and neglects much delicate information. To better utilize those predictions and estimate the uncertainty, we further propose feature-perturbed consistency to exploit reliable regions for our framework to learn from. Extensive experiments on the public BraTS2020 dataset and the 2017 ACDC dataset confirm the efficiency and effectiveness of our method. In particular, the proposed method demonstrates remarkable superiority in the segmentation of boundary regions. The project is available at https://github.com/youngyzzZ/SFPC.**

Index Terms—**Semi-supervised learning, Uncertainty estimation, Image segmentation.**

I. Introduction

Automated semantic segmentation is a crucial and fundamental task in medical image analysis and state-of-the-art performance in various segmentation tasks has been achieved by fully supervised learning approaches [1]–[6]. However, fully supervised learning approaches require sufficient and precise annotations to train models for satisfactory performance. Acquiring such a large-scale dataset with pixel-wise annotation is often challenging because it is expensive and time-consuming. To overcome the annotation scarcity, a promising approach is to adopt semi-supervised learning, which typically utilizes a combination of a limited set of labeled samples and an adequate set of unlabeled ones for effective model training.

Considerable efforts have been devoted to reducing the annotation cost by efficiently leveraging unlabeled data to improve the segmentation performance in the semi-supervised community [7]–[11]. Existing semi-supervised methods can be broadly classified into three categories. The first category refers to those methods that generate pseudo labels for unlabeled images and consider them as ground-truth labels to leverage more information about unlabeled images [12]–[14]. The second category refers to those consistency-based methods [15]–[19] that impose small perturbations to inputs to obtain predictions with subtle differences. Those methods are based on the assumption that the predictions from the model should be consistent under different input perturbations. For instance, Yu *et al.* [17] introduced an uncertainty-aware method for left atrium image segmentation, where the teacher model generates more reliable labels for student models to learn from and simultaneously estimates the uncertainty of the label. Li *et al.* [20] and Wang *et al.* [21] further investigate the shape constraints via introducing the signed distance map and the signed distance field respectively. Wu *et al.* [22] designed a mutual consistency network to use the unannotated images by encouraging the predictions of three slightly different decoders to be consistent. With the remarkable improvement achieved by consistent regularization [23], [24] methods, Luo *et al.* [25]. proposed a pyramid-prediction network with uncertainty rectified pyramid consistency for lesion segmentation, but the predictions generated from the shallow layers tend to be coarse and imprecise. The third category refers to several powerful similarity learning approaches, such as contrastive learning, which have been employed in semi-supervised learning [26], [27], in which a classification model with powerful feature extraction capabilities is pretrained and then effectively transferred for a segmentation task. For example, Gu *et al.* [28] proposed a cross-domain contrastive learning strategy to encourage extracting domain invariant features, meanwhile introducing a self-ensembling mean-teacher framework to exploit unlabeled target domain images with a prediction consistency constraint. Hu *et al.* [29] designed a supervised local contrastive loss that leverages limited pixel-wise annotation to force pixels with the same label to gather around in the embedding space for better performance. Chaitanya *et al.* [30] presented a local contrastive loss to learn good pixel-level features useful for segmentation by exploiting semantic label information obtained from pseudo-

[*] Coresponding author
Email-to: sujingyong@hit.edu.cn

979-8-3503-3748-8/23/$31.00 © 2023 IEEE

labels. Nevertheless, methods that rely on contrastive learning necessitate meticulous design of upstream tasks and substantial volumes of data for training an effective feature extractor.

Most existing consistency-based methods generate multiple predictions relying on network architectures such as multiple decoders, pyramid structure or MC dropout, without considering that perturbations at the feature level could result in disparate model outputs. In this work, we first propose a novel semi-supervised framework with a feature-level perturbation for medical image segmentation, aiming to leverage more information from unlabeled data via uncertainty estimation. The proposed network is composed of one shared encoder and two slightly different decoders. Specifically, a semantic-level perturbation module is integrated into the segmentation decoder, which enables our model to generate a batch of predictions with slight differences. Then, the statistical discrepancy of predictions for a certain input is used to estimate the pixel/voxel-level uncertainty. In addition, a reconstruction task is introduced to help the segmentation branch contain and capture more structural information. We also design a new feature-perturbed consistency training scheme. Our main contributions are summarized as follows:

- We propose a novel Feature-Level Perturbation Module (FLPM) to explore the abundant useful information of unlabelled medical images.
- We design a feature-perturbed consistency to emphasize reliable predictions and weaken unreliable ones for efficient training.
- Extensive evaluations on both two-dimensional and three-dimensional datasets demonstrate the efficacy of our method, with new state-of-the-art performance achieved in semi-supervised segmentation, especially in the boundary regions.

II. Method

This study aims to develop a general semi-supervised learning framework which can utilize plenty of unlabelled training data to help train a three- or two-dimensional segmentation model. Here a novel feature-level perturbation strategy is proposed to help the model effectively utilize unlabelled training data during semi-supervised learning.

A. The overall semi-supervised learning framework

An overview of the proposed semi-supervised learning (SSL) framework is illustrated in Figure 1. The architecture of the framework contains an encoder E, a decoder G_1 for segmentation, and an auxiliary decoder G_2 for reconstruction. Both decoders share the same encoder E. Specifically, a semantic-level perturbation module is included in the segmentation decoder G_1 for unsupervised data during model training (Figure 1). With multiple times of feature perturbations and the corresponding segmentation prediction outputs for each unlabelled input image, an innovative semantic-level consistency loss for unsupervised data is designed by enforcing the feature-perturbed prediction results to be consistent. Such consistency loss is expected to help train a more robust segmentation model. On the other hand, the auxiliary reconstruction task (Figure 1, lower right) is expected to help the shared encoder E achieve more powerful encoding capability which in turn would benefit the segmentation task.

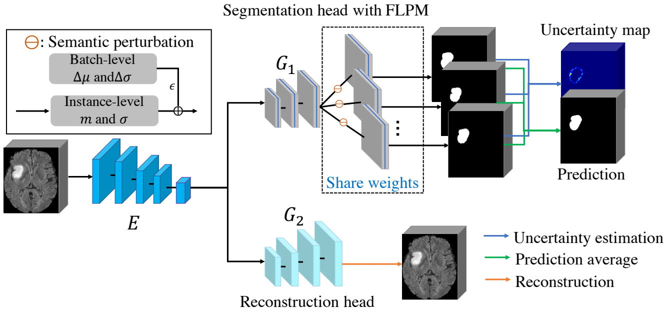

Fig. 1: Overview of the feature-perturbed semi-supervised segmentation framework. Two complementary tasks are performed. The feature-level perturbation module (FLPM) is seamlessly integrated into the segmentation decoder. Skip connection between each stage of the encoder and the corresponding stage of each decoder is omitted for clarity.

B. Feature-level perturbation

The idea of perturbing features is inspired by the recently proposed method MaxStyle [31] which shows that different linear transformations of the same feature maps may generate synthetic images with different degrees of contrast. Different from MaxStyle which is developed to generate various styles of synthetic data from the auxiliary decoder for improving model's out-of-domain (OOD) robustness, the proposed feature perturbation here is performed at one top-level layer of the segmentation decoder to help the model effectively utilize unlabelled data in semi-supervised learning (Figure 2). Formally, for each unlabelled training data $\mathbf{x}_j$, denote by $\mathbf{f}_{j,c}$ the corresponding c-th channel of the feature map output from a pre-determined (e.g., second last convolutional) layer in the decoder G_1, and $m(\mathbf{f}_{j,c})$ and $\sigma(\mathbf{f}_{j,c})$ respectively the mean and standard deviation of all elements in $\mathbf{f}_{j,c}$. Then, $\mathbf{f}_{j,c}$ can be randomly perturbed by a linear transformation of its normalized version $\bar{\mathbf{f}}_{j,c} = \frac{\mathbf{f}_{j,c} - m(\mathbf{f}_{j,c})}{\sigma(\mathbf{f}_{j,c})}$ as below,

$$\hat{\mathbf{f}}_{j,c} = \gamma_{j,c} \cdot \bar{\mathbf{f}}_{j,c} + \boldsymbol{\mu}_{j,c} \tag{1}$$

$$\gamma_{j,c} = \sigma(\mathbf{f}_{j,c}) + \epsilon \cdot \Delta\sigma_c \tag{2}$$

$$\boldsymbol{\mu}_{j,c} = m(\mathbf{f}_{j,c}) + \epsilon \cdot \Delta\mu_c \tag{3}$$

where γ_j and $\boldsymbol{\mu}_{j,c}$ are slightly perturbed version of the standard deviation $\sigma(\mathbf{f}_{j,c})$ and the mean $m(\mathbf{f}_{j,c})$, respectively. The perturbations are controlled respectively by the estimated standard deviations ($\Delta\sigma_c$ and $\Delta\mu_c$) of $\sigma(\mathbf{f}_{j,c})$ and $m(\mathbf{f}_{j,c})$ over a mini-batch of unlabelled training images (including $\mathbf{x}_j$) during model training, and by a randomly sampled scale ϵ from a uniform distribution within the range $[-\tau, \tau]$. $\Delta\sigma_c$ and $\Delta\mu_c$ can help control the perturbations within a reasonable range such that any perturbed feature map $\hat{\mathbf{f}}_{j,c}$ is semantically

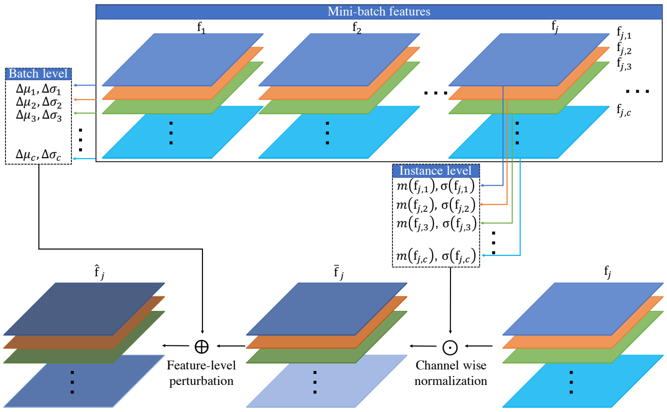
Mini-batch features
f_1
f_2
f_j
$f_{j,1}$
$f_{j,2}$
$f_{j,3}$
$f_{j,c}$
Batch level
$\Delta\mu_1, \Delta\sigma_1$
$\Delta\mu_2, \Delta\sigma_2$
$\Delta\mu_3, \Delta\sigma_3$
$\Delta\mu_c, \Delta\sigma_c$
Instance level
$m(f_{j,1}), \sigma(f_{j,1})$
$m(f_{j,2}), \sigma(f_{j,2})$
$m(f_{j,3}), \sigma(f_{j,3})$
$m(f_{j,c}), \sigma(f_{j,c})$
$\hat{f}_j$
$\bar{f}_j$
f_j
$\oplus$
Feature-level perturbation
$\odot$
Channel wise normalization

Fig. 2: The Feature-Level Perturbation Module (FLPM). Means and standard deviations at both the instance and batch levels are applied to each feature map at a top convolutional layer for each image within a mini-batch.

meaningful. Note that the same scale variable ϵ is used across all feature channels ($\{c\}$) such that all channels of feature maps have the same level of perturbation. The perturbed feature maps are finally fed to the subsequent convolutional layer(s), resulting in the segmentation probability output for the specific feature-level perturbation of the input $\mathbf{x}_j$.

C. Feature-perturbed consistency for semi-supervised learning

In semi-supervised learning, one widely used strategy is to design a consistency loss on unlabelled training data. The basic idea is to enforce the model to generate similar outputs for two or more transformed versions of the same input. The perturbed feature maps and associated segmentation output probabilities here provide a natural way to design the consistency loss. In detail, let $D_u = \{\mathbf{x}_j, j = 1, \ldots, J\}$ denote the unlabelled training set containing J unlabelled images and each image contains K elements (pixels or voxels). For the j-th training unlabelled data $\mathbf{x}_j \in D_u$, suppose M independent perturbations are performed at the feature level as described above, resulting in M output probability maps from the segmentation decoder G_1. Let $\{\mathbf{p}_{j,k,m}, m = 1, \ldots, M\}$ represent the M output probability vectors for the k-th element of the input $\mathbf{x}_j$. The consistency loss based on feature perturbations can be represented as

$$\mathcal{L}_u = \frac{1}{J \cdot K} \sum_{j=1}^{J} \sum_{k=1}^{K} g(\mathbf{p}_{j,k,1}, \ldots, \mathbf{p}_{j,k,M}) \,, \tag{4}$$

where the consistency measurement $g(\cdot)$ can be any reasonable function measuring the consistency (often similarity) between all the M outputs $\{\mathbf{p}_{j,k,m}, m = 1, \ldots, M\}$ for each image element. Inspired by the recent work [25], we designed the consistency measurement function as

$$g(\mathbf{p}_{j,k,1}, \ldots, \mathbf{p}_{j,k,M}) = \frac{1}{M} \sum_{m=1}^{M} \frac{\omega_{j,k,m} \|\mathbf{p}_{j,k,m} - \bar{\mathbf{p}}_{j,k}\|}{\sum_{m=1}^{M} \omega_{j,k,m}} \,, \tag{5}$$

where $\bar{\mathbf{p}}_{j,k}$ is the average probability vector over all the M vectors $\{\mathbf{p}_{j,k,m}, m = 1, \ldots, M\}$, and $\omega_{j,k,m} = \exp\{-h(\mathbf{p}_{j,k,m})\}$ with $h(\mathbf{p}_{j,k,m})$ being the entropy of the discrete probability $\mathbf{p}_{j,k,m}$. The term $\|\mathbf{p}_{j,k,m} - \bar{\mathbf{p}}_{j,k}\|$ represents the L_p norm ($p = 1$ or 2) of the difference between a single output prediction $\mathbf{p}_{j,k,m}$ and the averaged prediction $\bar{\mathbf{p}}_{j,k}$. Therefore, minimizing this term would enforce the model to have similar predictions for all the M perturbed features from the same input data. On the other hand, entropy is a measurement of prediction uncertainty, and larger entropy would lead to smaller weight $\omega_{j,k,m}$. This will help the consistency measurement function pay more attention to confident predictions rather than unconfident (*i.e.*, uncertain) ones, and enforce that the confident predictions should be consistent. This is reasonable because there are often unconfident predictions around the boundary of regions.

The overall semi-supervised learning loss function can be then designed as

$$\mathcal{L} = \mathcal{L}_s + \lambda_1 \mathcal{L}_u + \lambda_2 \mathcal{L}_r \,, \tag{6}$$

where $\mathcal{L}_s$ is supervised loss (e.g., cross-entropy loss) only on the labeled training set, and $\mathcal{L}_r$ is the reconstruction loss (with L_2 norm) from the decoder G_2 on both the labeled and unlabelled sets. λ_1 and λ_2 are two coefficients to balance the there loss terms. Once the model is well trained, the encoder E and the segmentation decoder G_1 are used as a UNet model for segmentation of any new image during inference, where feature perturbation is not necessary.

III. Experiment

A. Datasets and evaluation metrics

In this study, we evaluate our method on the public datasets, BraTS2020 for whole brain tumor segmentation and 2017 ACDC for cardiac segmentation. The BraTS2020 [32] dataset contains 496 subjects, where 380, 26 and 90 subjects are assigned scans for training, validation and testing respectively. Note that the T2-FLAIR modality with isotropic $1mm^3$ resolution is adopted for our experiments. Each instance is normalized by its channel-wise means and standard deviations. The 2017 ACDC [33] dataset has 100 subjects, from which 75, 5 and 20 subjects are randomly selected for training, validating and testing, respectively. The intensity of each scan is re-scaled to $[0, 1]$. For semi-supervised partitions, 10% or 20% of training images were randomly selected as labeled samples and the remaining ones as unlabeled. To quantitatively assess the performance, three common evaluation metrics are adopted, i.e., Dice Similarity Coefficient (DSC), 95% Hausdorff Distance (95HD) and the Average Surface Distance (ASD).

B. Implementation details

Our framework is implemented by PyTorch and trained with two NVIDIA GeForce 3090 GPUs with 24 GB memory. The whole neural network is updated by an SGD optimizer (weight decay 1e-4, momentum 0.9) for 6000 iterations, with an initial learning rate of 0.01 decayed by 0.1 every 2500 iterations. The batch size was 8, consisting of 4 labeled images and 4 unlabeled images. We randomly cropped $112 \times 112 \times 112$ sub-volumes as

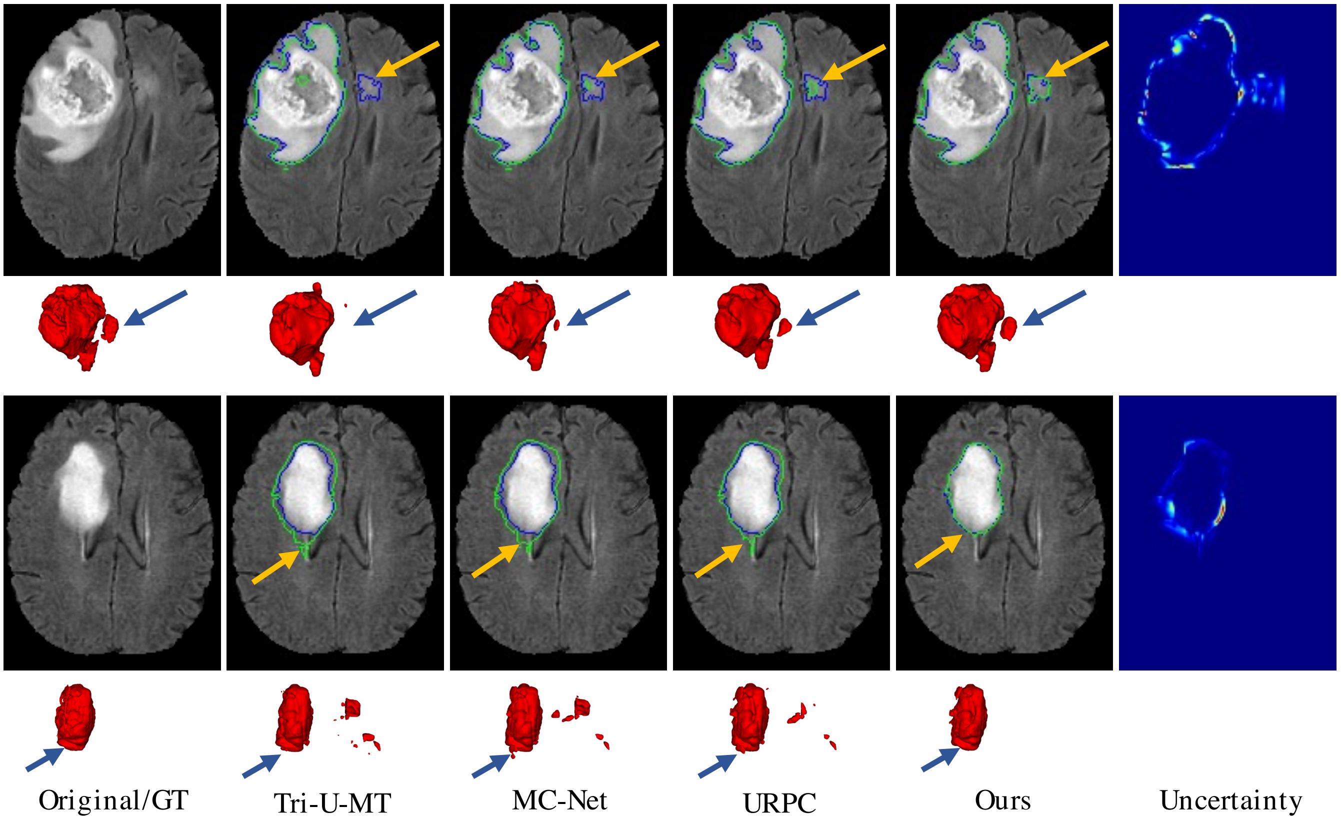

Fig. 3: Visual comparisons between the proposed method (last two columns) and strong baseline methods (second to fourth column) on two representative images from BraTS2020. During training, 10% training samples were annotated. Green and blue contours denote boundaries of the predicted and the ground-truth lesion regions, respectively. Second and fourth row (in red): view of the segmented 3D lesions. Last column: the prediction uncertainty for each pixel using entropy.

the network input for 3D volumes and resized 256×256 as the input for 2D slices. Data augmentation was employed to avoid over-fitting, containing random cropping, flipping and rotation. The final segmentation results of 3D volumes were obtained utilizing a sliding window strategy. For hyper-parameter setting, M was set to 4, and τ was set to 0.1 for subtle perturbations at semantic level. λ_1 was an elaborately designed time-dependent Gaussian warming up function [17], [34] to balance the weight between supervised and unsupervised learning stably, which was defined as $\lambda(t) = \omega_{max} \cdot e^{-5(1-\frac{t}{t_{max}})^2}$, where ω_{max} denotes the final regularization weight, t is the current training round and t_{max} denotes the maximal training round. According to the relevant study [17], ω_{max} was set to 0.1 for all experiments. λ_2 was simply set to 0.5.

C. Comparison with other semi-supervised methods.

To demonstrate the superiority of the proposed method in semi-supervised learning, we compared the lesion segmentation ability of our method with several state-of-the-art methods including nnU-Net [35], SASSnet [20], UAMT [17], Tri-UMT [21], DTC [15], CoraNet [24], MC-Net [22], PLCT [30], and URPC [25]. It is worth noting that only nnU-Net is trained in a fully supervised manner as the performance upper bound. Table I shows the quantitative results on BraTS2020 dataset with different labeled sample proportions of the training set. As observed, our proposed method outperforms all the compared semi-supervised methods with the highest DSC (84.83% and 86.35%) and lowest 95HD (10.79 and 8.64) respectively in two different settings. Compared with the strongest baseline URPC, the absolute improvement by the proposed method is respectively 0.75% and 0.74% in DSC, 0.77 and 0.27 in 95HD and 0.21 and 0.18 in ASD. Besides, when less labeled samples are utilized during training, our method leads to a higher superiority, demonstrating our method effectively leverages the unlabeled scans for performance gains. As illustrated in Figure 3, our method (fifth column) can more accurately locate the ambiguous regions (pointed by yellow and blue arrows in 2D and 3D views, respectively) and segment edge regions more accurately compared with other baselines (second to fourth column) on the BraTS2020 dataset. The pixel-level prediction uncertainty from our method (last column) can well indicate the challenging regions for segmentation, from which we can see the uncertain regions are mainly around lesion boundaries. Similar results were obtained on 2017 ACDC dataset. As Table I (right half) shows, our method outperforms all the strong semi-supervised baselines on all the three metrics. Figure 4

TABLE I: Quantitative comparisons with other state-of-the-art methods on BraTS2020 and 2017 ACDC datasets. ↑ indicates that larger values are better and ↓ indicates that smaller values are better.

Method	% scans used		BraTS2020 (3D)			2017 ACDC (2D)		
	Labeld	Unlabeld	DSC(%) ↑	95HD(mm) ↓	ASD(mm) ↓	DSC(%) ↑	95HD(mm) ↓	ASD(mm) ↓
SASSNet [20]	10	90	82.16	14.86	4.15	84.26	6.08	1.76
UAMT [17]			80.88	17.63	6.86	81.32	13.17	3.77
Tri-U-MT [21]			82.70	15.26	3.62	83.71	7.54	2.73
DTC [15]			81.86	16.31	3.67	82.43	8.82	3.15
CoraNet [24]			81.29	13.97	3.96	84.17	6.18	2.41
MC-Net [22]			83.75	13.55	3.34	86.55	7.01	2.13
PLCT [30]			83.51	13.74	3.62	86.48	6.69	2.32
URPC [25]			84.08	11.56	3.28	84.72	5.12	1.64
Ours			**84.83**	**10.79**	**3.07**	**87.52**	**4.96**	**1.33**
SASSNet [20]	20	80	84.67	9.41	2.64	86.98	5.36	2.52
UAMT [17]			84.86	12.21	2.19	85.62	9.31	1.53
Tri-U-MT [21]			85.02	8.83	3.16	87.04	5.62	1.63
DTC [15]			84.82	12.69	3.43	86.13	6.28	2.31
CoraNet [24]			84.37	9.05	2.62	86.32	6.45	2.21
MC-Net [22]			85.17	9.72	3.01	88.35	5.76	1.92
PLCT [30]			85.38	8.72	2.94	88.26	5.84	2.11
URPC [25]			85.61	8.91	2.55	87.18	5.29	1.61
Ours			**86.35**	**8.64**	**2.37**	**89.13**	**5.09**	**1.48**
nn-UNet	100	0	89.29	7.97	1.83	92.12	1.73	0.52

demonstrates two representative segmentation results from our method (sixth column) and the strong baselines (third to fifth column), again confirming the more accurate segmentation performance particularly around region boundaries.

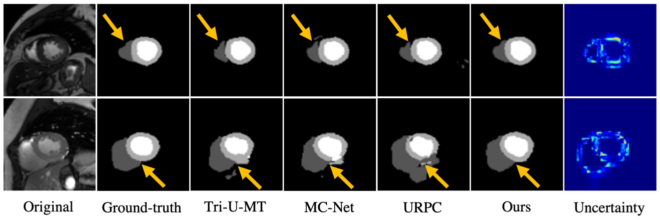

Fig. 4: Visual comparisons between the proposed method (last two columns) and strong baselines (third to fifth column) on two representative images from 2017 ACDC dataset. 10% training samples were annotated for model training. Second to sixth column: different gray values denote different types of segmented regions.

It is worth noting that the seemingly small difference in dice score (DSC) between our method and the strong baselines are probably because DSC is a metric based on the whole segmentation regions, and current strong baselines already exhibit satisfactory performance in segmenting the main region of interest. However, it remains challenging for those methods to accurately segment the boundary regions. Actually, when evaluated only around the boundary regions, defined as the band with a width of 10 pixels through boundary pixel expansion, our method has a more significant performance gain, outperforming the best baseline PLCT by 12.57% and 15.38% in DSC respectively on the BraTS2020 and 2017 ACDC datasets (Figure 5).

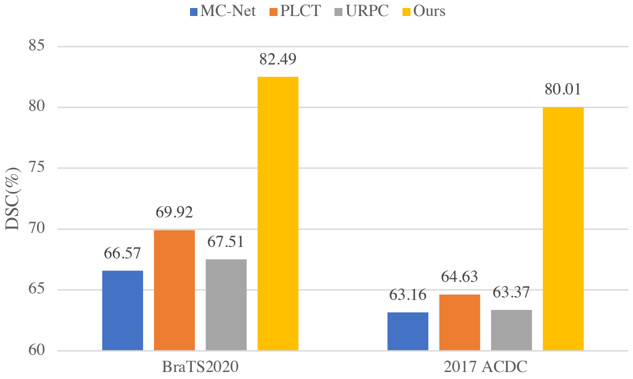

Fig. 5: Performance on boundary regions from our method and strong baselines on BraTS2020 and 2017 ACDC datasets, with 10% labeled images used in training.

D. Sensitivity study

Hyper-parameter M: M controls the number of predictions, which plays a vital role in stable training and uncertainty estimation. As demonstrated in Fig. 6, when varying the value of M in a large range (e.g., $[4, 8]$), the performance of our method is stably and better than that of the best baseline (dashed lines) on both datasets in all three metrics. This confirms that our method is insensitive to the choice of hyper-parameter M.

Hyper-parameter λ_2: While the coefficient λ_1 in the loss function is automatically adjusted over training iterations, the choice of the other coefficient λ_2 could affect the performance of our method. By varying value of λ_2 from 0.1 to 2.0, we observed that our method performs stably well within a large range $[0.5, 2]$, where our method always outperforms the strongest baseline (dashed lines) in both semi-supervised settings (Figure 7). The decreased performance with very small λ_2 values (e.g., 0.1) is probably due to the decreasing effect of the reconstruction decoder on performance boosting when λ_2 is too small.

Hyper-parameter τ: To achieve multiple meaningful segmentation results for each input image, our method applies a sensible perturbation to the feature layer, with the magnitude of the perturbation being adjusted by the hyper-parameter τ.

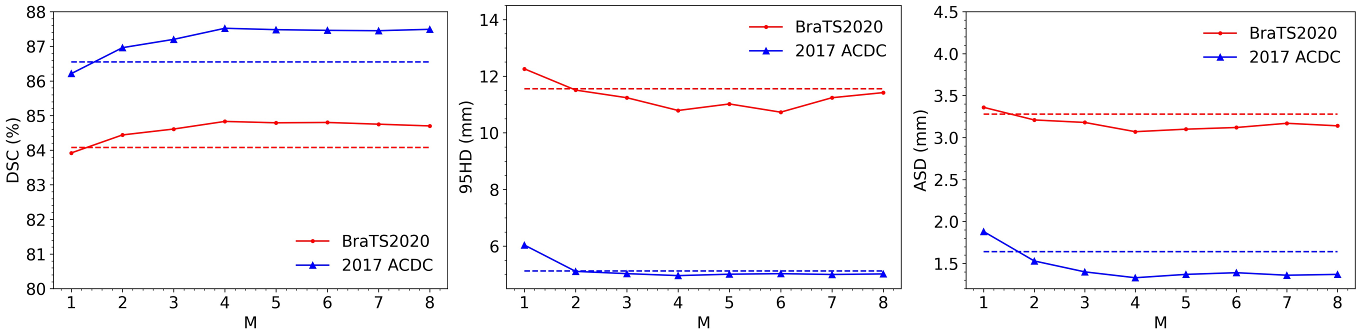

Fig. 6: Sensitivity analysis of hyper-parameter M in our method on BraTS2020 and 2017 ACDC datasets, where 10% labeled images were used for model training. Dashed lines represent the performance of the strongest baselines.

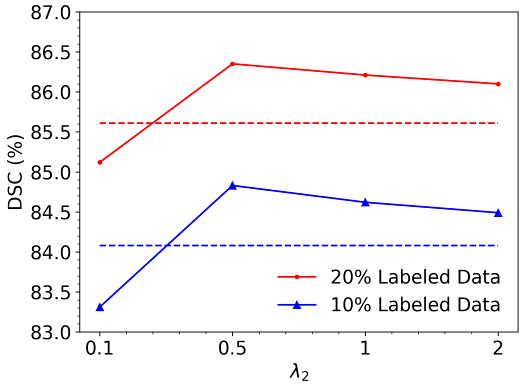

Fig. 7: Sensitivity analysis of the loss weight λ_2 in our method on the BraTS2020 dataset, where 10% and 20% labeled images were used for model training, respectively. Dashed lines represent performance of the strongest baseline URPC.

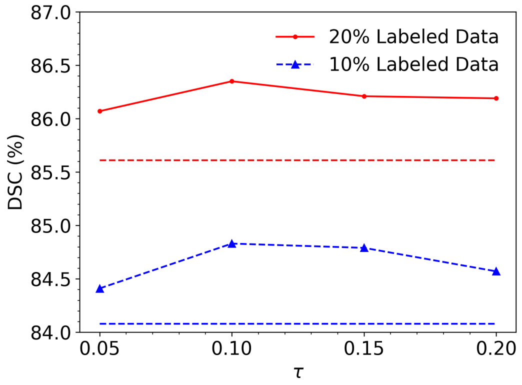

Fig. 8: Sensitivity analysis of τ in our method on the BraTS2020 dataset, where 10% and 20% labeled images were used in model training, respectively. Dashed lines represent performance of the strongest baseline URPC.

Fig. 8 shows the DSC performance of our method trained with different degrees of perturbation by τ on the BraTS2020 dataset. The results indicate that, in both semi-supervised settings, the DSC performance with different τ values changes little, suggesting that our method is robust to choice of the hyper-parameter τ. Note that too small or too large τ values lead to slightly worse performance, because a larger τ would cause excessive perturbation to the features of the middle layer, resulting in the loss of structural information and ultimately leading to inaccurate segmentation results, while a smaller τ may not introduce enough perturbation, resulting in overly consistent outputs from the decoders that hinder the model's ability to evaluate uncertainty and perform high-precision segmentation of edge regions.

E. Ablation study

To investigate the effect of different components in our framework, we conduct detailed ablation study on BraTS2020 and 2017 ACDC. As shown in Table II, the first row is a UNet trained with only labeled data (first row), which is the backbone of our framework. On the BraTS2020 dataset, as the reconstruction head was integrated into the network (second row), such joint learning improves segmentation results by 0.24% in DSC. We then employed the proposed feature-level perturbation module to the UNet (third row), which significantly improves the performance by 6.17% in DSC. While the reconstruction head and feature-level perturbation module are introduced into the framework simultaneously (fourth row), the DSC remarkably increases by 6.79% compared with the UNet. As the feature-perturbed consistency is adapted to the framework (last row), the DSC value reaches 84.83%. Similar performance improvement by each framework component was observed on the 2017 ACDC dataset (Table II, right half).

F. Effects of different perturbation methods

One crucial aspect of consistency training is the application of perturbations to the feature representation at certain hidden layer. To investigate the effects of adopted feature perturbation strategy in our method on model performance, three additional feature perturbation strategies, namely F-Noise [36], F-Drop [36], and Spatial Dropout [37] were used to replace the perturbation strategy in our method. Figure 9 demonstrates the segmentation performance based on different feature perturbation strategies. It is clear that the perturbation

TABLE II: Ablation study of our method on BraTS2020 and 2017 ACDC datasets, where 10% labeled images were used for model training. Seg, FPC, FLPM and Rec denote the segmentation head, feature-perturbed consistency, feature-level perturbation module and reconstruction head, respectively.

Seg	Rec	FLPM	FPC	**BraTS2020 (3D)**			**2017 ACDC (2D)**		
				DSC(%) ↑	95HD(mm) ↓	ASD(voxel) ↓	DSC(%) ↑	95HD(mm) ↓	ASD(voxel) ↓
✓				77.79	15.26	5.34	79.71	7.54	4.73
✓	✓			78.03	14.52	5.07	79.98	7.12	4.41
✓		✓		83.96	11.01	3.25	86.35	5.13	1.57
✓	✓	✓		84.58	10.92	3.16	87.15	5.06	1.42
✓		✓	✓	84.69	11.07	3.20	87.36	5.09	1.51
✓	✓	✓	✓	**84.83**	**10.79**	**3.07**	**87.52**	**4.96**	**1.33**

strategy in our method yields the highest DSC value in both semi-supervised learning settings. It is noteworthy that when employing the three alternative feature-level perturbation strategies, multiple forward computations are necessary in order to obtain multiple prediction results. Therefore, additional computational overhead is required in these strategies compared to the perturbation strategy in our method.

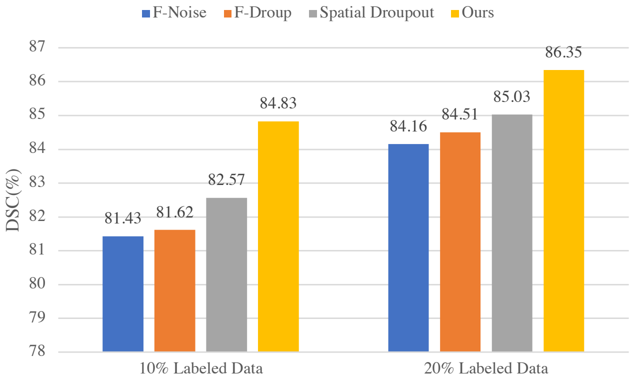

Fig. 9: Segmentation performance with different perturbation strategies on the BraTS2020 dataset, where 10% and 20% labeled images were used for model training, respectively.

G. Limitations and further work

Despite the superior performance of our framework in the context of semi-supervised medical image segmentation, the current perturbation strategy, which is based on mini-batch images, has the potential to be extended to encompass the entire dataset. There is also merit in exploring a wider range of model architectures for novel tasks, as this has the potential to enhance the generation of meaningful semantic perturbations within the model. Besides, here we only investigate the feature-level perturbations, while the conventional data-level perturbations could also be effective and combined together with the feature-level perturbations. Other future work includes investigating more efficient feature-level perturbation strategies for adapting different segmentation tasks and validating the proposed framework with larger and more diverse segmentation tasks.

IV. Conclusion

In this paper, we propose a novel and efficient semi-supervised framework with feature-level perturbations for medical image segmentation. In contrast to previous methods, our method leverages more information from unlabeled images at the feature level to facilitate effective segmentation and promote stable network learning. Based on a dual-task architecture, our reconstruction network enforces the encoder to maintain more semantic information for the segmentation branch to perform stably. To efficiently leverage the abundant information naturally inherited in the unlabeled data, we employ a feature-level perturbation module. A feature-perturbed consistency is introduced to leverage reliable information from unlabeled images and further improve the model performance. Extensive experimental results demonstrate the feasibility and superiority of our framework, especially around the boundary regions of lesions. The applications of our framework to more medical image segmentation tasks will be explored in future work.

Data Use Declaration: Our experimental data were collected from open source datasets. The BraTS2020 can be downloaded at: https://www.med.upenn.edu/cbica/brats2020/data.html, and the 2017 ACDC dataset is available at: https://humanheart-project.creatis.insa-lyon.fr/database/#

Acknowledgments

This work is supported by National Natural Science Foundation of China (grant No. 62376068), by the Guangdong International Collaborative Project (grant No. 2023A0505050109), by the Shenzhen Science and Technology Innovation Program (grant No. JCYJ20220818102414031), by the Major Key Project of PCL (grant No. PCL2023AS7-1), by the National Natural Science Foundation of China (grant No. 62071502), and Guangdong Excellent Youth Team Program (grant No. 2023B1515040025).

References

[1] Ziyun Yang and Sina Farsiu, "Directional connectivity-based segmentation of medical images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11525–11535.

[2] Chenyang Lu, Daan de Geus, and Gijs Dubbelman, "Content-aware token sharing for efficient semantic segmentation with vision transformers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23631–23640.

[3] Saumya Gupta, Xiaoling Hu, James Kaan, Michael Jin, Mutshipay Mpoy, Katherine Chung, Gagandeep Singh, Mary Saltz, Tahsin Kurc, Joel Saltz, et al., "Learning topological interactions for multi-class medical image segmentation," in *European Conference on Computer Vision*, 2022, pp. 701–718.

[4] Ye Huang, Di Kang, Liang Chen, Xuefei Zhe, Wenjing Jia, Linchao Bao, and Xiangjian He, "Car: Class-aware regularizations for semantic segmentation," in *European Conference on Computer Vision*, 2022, pp. 518–534.

[5] Gongping Chen, Lei Li, Yu Dai, Jianxun Zhang, and Moi Hoon Yap, "Aau-net: an adaptive attention u-net for breast lesions segmentation in ultrasound images," *IEEE Transactions on Medical Imaging*, vol. 42, no. 5, pp. 1289–1300, 2023.

[6] Jia-Xin Zhuang, Jiabin Cai, Jianguo Zhang, Wei-shi Zheng, and Ruixuan Wang, "Class attention to regions of lesion for imbalanced medical image recognition," *Neurocomputing*, vol. 555, pp. 126577, 2023.

[7] Hoel Kervadec, Jose Dolz, Éric Granger, and Ismail Ben Ayed, "Curriculum semi-supervised segmentation," in *Medical Image Computing and Computer Assisted Intervention*, 2019, pp. 568–576.

[8] Gerda Bortsova, Florian Dubost, Laurens Hogeweg, Ioannis Katramados, and Marleen De Bruijne, "Semi-supervised medical image segmentation via learning consistency under transformations," in *Medical Image Computing and Computer Assisted Intervention*, 2019, pp. 810–818.

[9] Wenlong Hang, Wei Feng, Shuang Liang, Lequan Yu, Qiong Wang, Kup-Sze Choi, and Jing Qin, "Local and global structure-aware entropy regularized mean teacher model for 3D left atrium segmentation," in *Medical Image Computing and Computer Assisted Intervention*, 2020, pp. 562–571.

[10] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng, "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 523–534, 2020.

[11] Ruiguo Yu, Xianzhi Zhang, Mankun Zhao, Yang Yan, Ming Li, and Mei Yu, "Caanet: Cam-guided adaptive attention network for weakly supervised semantic segmentation of thyroid nodules," in *IEEE International Conference on Bioinformatics and Biomedicine*, 2022, pp. 1791–1795.

[12] Hao Zheng, Susan M Motch Perrine, M Kathleen Pitirri, Kazuhiko Kawasaki, Chaoli Wang, Joan T Richtsmeier, and Danny Z Chen, "Cartilage segmentation in high-resolution 3D Micro-CT images via uncertainty-guided self-training with very sparse annotation," in *Medical Image Computing and Computer Assisted Intervention*, 2020, pp. 802–812.

[13] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S. Rawat, and Mubarak Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," in *International Conference on Learning Representations*, 2021.

[14] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2613–2622.

[15] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang, "Semi-supervised medical image segmentation through dual-task consistency," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, number 10, pp. 8801–8809.

[16] Xiangde Luo, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Nianyong Chen, Guotai Wang, and Shaoting Zhang, "Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency," in *Medical Image Computing and Computer Assisted Intervention*, 2021, pp. 318–329.

[17] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *Medical Image Computing and Computer Assisted Intervention*, 2019, pp. 605–613.

[18] Dafei Qiu, Jiajin Yi, and Jialin Peng, "Wda-net: Weakly-supervised domain adaptive segmentation of electron microscopy," in *IEEE International Conference on Bioinformatics and Biomedicine*, 2022, pp. 1132–1137.

[19] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12546–12558, 2020.

[20] Shuailin Li, Chuyu Zhang, and Xuming He, "Shape-aware semi-supervised 3D semantic segmentation for medical images," in *Medical Image Computing and Computer Assisted Intervention*, 2020, pp. 552–561.

[21] Kaiping Wang, Bo Zhan, Chen Zu, Xi Wu, Jiliu Zhou, Luping Zhou, and Yan Wang, "Tripled-uncertainty guided mean teacher model for semi-supervised medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention*, 2021, pp. 450–460.

[22] Yicheng Wu, Zongyuan Ge, Donghao Zhang, Minfeng Xu, Lei Zhang, Yong Xia, and Jianfei Cai, "Mutual consistency learning for semi-supervised medical image segmentation," *Medical Image Analysis*, vol. 81, pp. 102530, 2022.

[23] Yixin Wang, Yao Zhang, Jiang Tian, Cheng Zhong, Zhongchao Shi, Yang Zhang, and Zhiqiang He, "Double-uncertainty weighted method for semi-supervised learning," in *Medical Image Computing and Computer Assisted Intervention*, 2020, pp. 542–551.

[24] Yinghuan Shi, Jian Zhang, Tong Ling, Jiwen Lu, Yefeng Zheng, Qian Yu, Lei Qi, and Yang Gao, "Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 41, no. 3, pp. 608–620, 2021.

[25] Xiangde Luo, Guotai Wang, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Dimitris N Metaxas, and Shaoting Zhang, "Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency," *Medical Image Analysis*, vol. 80, pp. 102517, 2022.

[26] Yuhang Zhang, Xiaopeng Zhang, Jie Li, Robert Qiu, Haohang Xu, and Qi Tian, "Semi-supervised contrastive learning with similarity co-calibration," *IEEE Transactions on Multimedia*, vol. 25, pp. 1749–1759, 2023.

[27] Hritam Basak and Zhaozheng Yin, "Pseudo-label guided contrastive learning for semi-supervised medical image segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19786–19797.

[28] Ran Gu, Jingyang Zhang, Guotai Wang, Wenhui Lei, Tao Song, Xiaofan Zhang, Kang Li, and Shaoting Zhang, "Contrastive semi-supervised learning for domain adaptive segmentation across similar anatomical structures," *IEEE Transactions on Medical Imaging*, vol. 42, no. 1, pp. 245–256, 2022.

[29] Xinrong Hu, Dewen Zeng, Xiaowei Xu, and Yiyu Shi, "Semi-supervised contrastive learning for label-efficient medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention*, 2021, pp. 481–490.

[30] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu, "Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation," *Medical Image Analysis*, vol. 87, pp. 102792, 2023.

[31] Chen Chen, Zeju Li, Cheng Ouyang, Matthew Sinclair, Wenjia Bai, and Daniel Rueckert, "Maxstyle: Adversarial style composition for robust medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention*, 2022, pp. 151–161.

[32] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[33] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al., "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?," *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[34] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[35] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[36] Yassine Ouali, Céline Hudelot, and Myriam Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12674–12684.

[37] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler, "Efficient object localization using convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 648–656.