Metric Learning with Feature Embedding for Segmentation Quality Evaluation

Huixiang Chen, Bo Peng, Zaid Al-Huda and Yifei Li

September 26, 2021

Metric Learning with Feature Embedding for Segmentation Quality Evaluation

1st H. Chen
School of Computing and Artificial Intelligence
Southwest Jiaotong University
Chengdu, Sichuan, China
chx880324@163.com

2nd B. Peng
School of Computing and Artificial Intelligence
Southwest Jiaotong University
Chengdu, Sichuan, China
bpeng@swjtu.edu.cn

3rd Z. Al-Huda
School of Computing and Artificial Intelligence
Southwest Jiaotong University
Chengdu, Sichuan, China
zaid.alhuda@my.swjtu.edu.cn

4th F. Li
School of Computing and Artificial Intelligence
Southwest Jiaotong University
Chengdu, Sichuan, China
xfLi@my.swjtu.edu.cn

***Abstract*—Segmentation quality evaluation is an essential step to quantify the performance of segmentation algorithms. It can be used as a feedback for correcting segmentation errors or selecting appropriate algorithm parameters. We propose a novel evaluation framework where a convolutional neural network is designed for distinguishing the segmentation quality instead of using ground truth images. Our work has three primary contributions: First, we evaluate the quality of object segmentation by learning region features. A novel feature embedding model is proposed to integrate meta evaluation principles in a metric learning process. Second, it exempts the requirement of ground truths in the test stage, where object features of trained classes are used for the discrepancy calculation. Third, a large-scale object segmentation evaluation dataset is constructed, which contains various segmentation qualities under different assumptions. The experimental results on PASCAL VOC2012 dataset demonstrate that our method improves the evaluation accuracy and outperforms the popular supervised evaluation measures.**

***Index Terms*—Image segmentation evaluation, metric learning, meta-measures, feature embedding, object segmentaion**

I. Introduction

Image segmentation is a necessary pre-processing step in computer vision tasks. In recent years, researchers have proposed many segmentation algorithms with high performance, while few work has been contributed to the study of segmentation quality evaluation. It is widely accepted that none of the existing segmentation algorithms is universally applicable to all images or scenarios, which makes a challenge for designing the evaluation methods. Usually, image objects consist of visually salient pixels and have definite semantic meanings. Therefore, evaluating the quality of object segmentation is a relatively objective task. In other words, people can easily agree on a common standard for evaluating the segmented objects in an image. Existing evaluation methods are mainly classified into three categories: analytical methods, empirical goodness methods and empirical discrepancy methods [1]. The analytical methods directly evaluate the principles, requirements and complexity of the segmentation algorithm itself, thus could be difficult for practical applications. On the other hand, empirical goodness methods based on human vision are proposed, which are relatively subjective and cannot be standardized. However, empirical discrepancy [26], [31], [19] is the most commonly used strategy. The basic idea is to compare the differences between a segmentation result and its human labeled counterpart (ground truth). To this end, various region features and edges [31], [9], [8], [29], [18] of

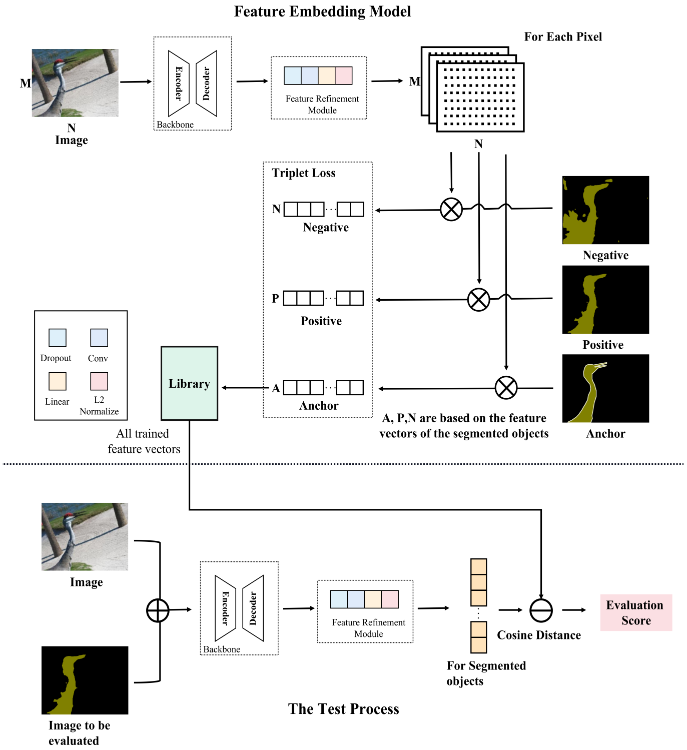

Fig. 1. Overview of the proposed framework.

the segmentation have been explored for designing measures in terms of distance or similarity. These methods only compute geometric features of object elements, i.e. region or edges, while image cues of the objects are not directly considered. In other words, object features such as colors, textures and semantic cues are not compared and are treated equally in the evaluation. Moreover, hand-crafted features mainly express low-level image information yet insufficient to represent the high-level semantic information.

In this paper, we study the segmentation evaluation problem in a learning based aspect. Following the principle of discrepancy evaluation, we design a convolutional neural network (CNN) for segmentation feature embedding, in the meanwhile integrating the meta-measures [22] into the metric learning process to extract distinguishable object features. Inspired by the object classification and segmentation tasks [25], [16], we pursue an evaluation method by learning to evaluate the segmentation quality instead of using pre-designed criteria. The proposed segmentation evaluation will be conducted in two steps. Firstly, an original image and its segmentation are passed into a Feature Embedding Model (FEM) to obtain the object feature vectors of the segmentation. Then, the cosine distance between these feature vectors and those from trained object classes is calculated as the final evaluation score.

To the best of our knowledge, it is the first attempt to explicitly learn a CNN for extracting segmentation features and couple the evaluation principles in a unified architecture.

The remaining sections of this paper are organized as follows. Section 2 explains related work. Section 3 presents he detailed proposed method. Section 4 provides the experimental results. Section 5 concludes the paper.

II. RELATED WORK

In recent years, there has been a growing interest in exploring segmentation quality evaluation using empirical discrepancy methods, since it has a generality in various segmentation algorithms, and quantitatively and objectively evaluate the quality of segmentation results. Researchers computed the discrepancy between handcrafted features, such as regions [4], [27], edges [3] or a combination [19]. For example, [4] represented region features that used global consistency errors and local consistency errors to calculate the accuracy of segmentation. In contrast, [3] explored edge features to evaluate segmentation quality. Movahedi et al. [19] combined region features and edge features to further enhance the efficiency of the evaluation. These methods take the assumption that the segmentation quality is only depended on the consistency between the used geometric features and the ground truths.

Researchers [13] [28] adopted CNN frameworks to design segmentation quality evaluation algorithms, taking into consideration hand-crafted features can only express low-level image information and the excellent performance of convolutional neural work in extracting image high-level semantic features. Shi et al. [28] proposed a double and multi scale deep CNN evaluation model, which seeks to obtain more comprehensive local and global information. The segmentation quality evaluation can also be viewed as a regression problem [13] and suggested three evaluation models, where an object detection network was proposed to predict the quality scores. The first is to modify the last layer of the network to a sigmoid layer. The second is to measure IoU for both the segmented image and the image produced by a specific segmentation algorithm. The third is to divide the segmented image into the foreground and background. But in principle, there is no clear explanation for the connection between the regression score and the similarity to the human labelings.

TABLE I
THE MAIN PARAMETERS OF THE THREE OBJECT SEGMENTATION ALGORITHMS.

	DeepLab v3	**FCN**	**Mask R-CNN**
input size	512	any size	any size
epoch	25	25	25
batch size	16	8	16
learning rate	0.007	1.0e-10	0.001
weight decay	0.0005	0.0005	0.0005
momentum	0.9	0.99	0.9

In general, the main principle in metric learning [33] is to shorten the distance between ground truth and positive samples and to expand the distance between ground truth and negative samples. Moreover, metric learning is commonly regarded as similarity learning [34]. When calculating the similarity between images, the purpose of metric learning to maximize the inter-class variations and minimize the intra-class variations. In order to deal with various feature similarities, in a particular task, we can pick appropriate features and manually construct a distance function. In a general sense, there are two types of metric learning [15]: metric learning by linear transformation and the nonlinear metric learning model.

There are evaluation methods studying the extent to which the segmentation matches the criteria of the good segmentations [3], [2], [20], [23], [32]. Common characteristics of objects (e.g. homogeneous regions, smooth boundaries, etc.) were explored, but they can not accurately quantify the complex objects in natural images [5]. Feature learning and integration [8] is a prominent way to obtain object features, however there is a gap between extracting features for regressing the segmentation quality and optimizing the metric value.

III. PROPOSED METHOD

The main contributions of this work are three-folds. First, to accurately describe objects features, a novel FEM is proposed, which is further integrated with the meta-measures for feature representation. Moreover, the score obtained by the proposed method is proportional to the effect of image segmentation. That is, the higher the score, the better the segmentation effect. Second, in the test stage, the trained object features are used for unsupervised evaluation. Compared to supervised evaluation methods, it exempts the requirement for ground truth, thus can be used for online evaluation. Third, we construct a large-scale object segmentation evaluation dataset, where different assumptions on the segmentation quality are used to obtain the positive and the negative samples.

A. Network architecture

The proposed CNN evaluation model is shown in Fig. 1. The training stage produces a set of library vectors, which represents object features from ground truths. Based on these vectors, the quality of arbitrary segmented objects can be evaluated in the test stage through the objects in the image. A FEM is designed to learn the feature embedding space, where a refinement step is performed on the feature space to get feature vector corresponding to the object region. Specifically, we use a U-Net structure [24] with a linear activation in the last layer to extract object features from the original image. Then a feature refinement module is used for calculating the embedding features for each image pixel. The module contains four layers, i.e. dropout, convolution, linear activation and L2 normalization. The ratio parameter in dropout is set as 0.5. In the convolutional layer, 3 × 3 × 3 convolution kernels are used with a linear activation, and the number of output channels is 3. The module finally outputs M × N × 3 embedding features. To extract object features, the segmentation mask is utilized, where the object pixels are set as ones, and background pixels as zeros. A dot multiplication is performed between each channel of the feature maps and the zero-one segmentation mask, so that feature values in the corresponding object regions are retained, while the non-object positions become zero. The product result is then reshaped to obtain a one-dimensional feature vector. In the training stage, we simultaneously use three segmentations of the same object, i.e., anchor (ground truth), positive and negative samples. The binary segmentation map is duplicated and converted into a N × M × 3 one-zero matrix, where N and M are the width and height of the image, respectively.

We seek to a network which can evaluate the segmentation quality under the principle of meta-measures. In [21], meta-measures were proposed to distinguish the qualities of different segmentations. Choosing an appropriate evaluation metrics (M), the meta-measure is defined as:

$$|M(S_1) - M(S_2)| < |M(S_1) - M(S_3)| \qquad (1)$$

where S_1 S_2 and S_3 are segmentations in different qualities, i.e., the ground truth, the positive sample, and the negative sample, respectively. In evaluation, the similarity between anchor and positive samples is expected to be larger than the similarities of anchor and negative samples. For the same image, the embedding object features do not directly denote the segmentation quality, however the similarity relations among the three entities will also follow the meta evaluation rule, as is shown in (2).

$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2 \qquad (2)$$

where for any object segmentation i, x_i^a is the anchor feature vector, x_i^p is the positive feature vector and x_i^n is the negative feature vector, α is a margin for enforcement between positive and negative pairs. Since the distance metric is established, we consider the similarity loss function as follows:

$$L = \sum^{i} \|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha - \|f(x_i^a) - f(x_i^n)\|_2^2 \qquad (3)$$

It is in the form of a triplet loss function. Using this similarity loss, we can learn the feature embedding space in an end-to-end manner. Finally, the feature vectors of all anchors are collected to form a feature library.

In the test stage, a segmentation mask is input with the source image. Then object segmentation features are calculated by FEM. We calculate the evaluation score as the cosine distance between the segmentation feature and the anchors' features in the library for the same object class. Because the proposed evaluation dataset is based on the VOC dataset, it contains 21 categories and category identifiers. When the feature vectors of extracted objects are stored in the library vectors, they will be mapped to the category identifiers one by one to ensure that each object has an accurate corresponding class identifiers. In the test, the feature vectors of target region in the original image are extracted and compared with the categories in the library vectors. The evaluation score for objects from the same class is defined as:

$$S(x_i) = max\left(\frac{x_i \cdot y}{|x_i|\,|y|}\right) \qquad (4)$$

where x_i,is the segmentation vector of the object, and $y \in Y$ is an anchor feature vector in the object set Y from the library. The highest score is taken as the final result. In presence of multiple object classes in an image, the average score is calculated by taking an equal importance of all objects, i.e.:

$$S = \frac{1}{n}\sum_{i=1}^{n} max\left(\frac{x_i \cdot y}{|x_i|\,|y|}\right) \qquad (5)$$

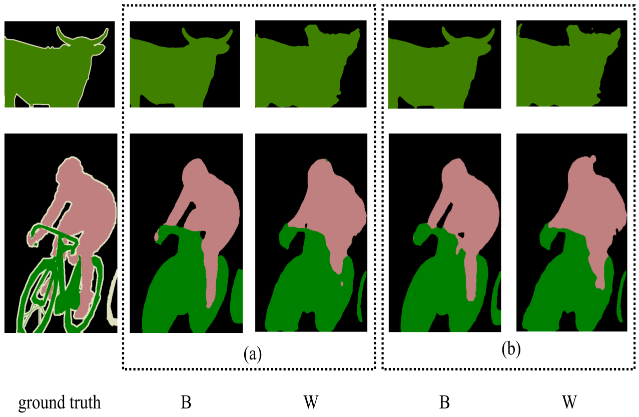

Fig. 2. Part of the dataset.From left to right: ground truth, B and W. a and b represent the 1st-rank test set and the 2nd-rank test set, respectively.

TABLE II
ACCURACY (%) OF UNSUPERVISED EVALUATION MEASURES.

measures	best positive	worst positive	1st-rank test set
F	44.2	38.5	54.6
$F^{'}$	51.9	42.3	62.6
Q	73.5	64.7	81.5
Ecw	43.3	37.5	63.3
Zeb	70.8	58.7	70.7
E	54.8	64.4	75.9
Ours	**76.3**	**72.9**	**90.1**

TABLE III
COMPARING THE ACCURACY (%) WITH SUPERVISED EVALUATION MEASURES.

measures	1st-rank test set	2nd-rank test set
Dice	77.372	71.167
PRI	79.562	77.737
SC	83.211	81.386
VI	83.576	83.211
IoU	**92.635**	90.032
Ours	91.793	**90.148**

B. Dataset

In the existing segmentation quality evaluation datasets, the segmentation images are independent, which does not represent the actual situation of segmented images [15]. Therefore, in order to evaluate the quality of segmentation images more reasonably and accurately, we should consider the correlations among segmentation images when constructing our dataset.

A segmentation evaluation database should be prepared for training the proposed deep evaluation network. We select images with 21 object classes from Pascal VOC 2012. Three popular object segmentation algorithms (i.e., DeepLab V3 [7], FCN [17], Mask R-CNN [12]) are used to create segmentation samples, where main parameters are set as in Table I. The segmentation results obtained by different CNNs can lead to various of qualities in the segmentations. Specifically, for DeepLab V3, we obtain the results from the 1st, 5th, 10th,15th and 25th epoches. And for FCN and Mask-R-CNN, we use the results from the 25th epoch. We collect 7 segmentations of each image for evaluation.

Firstly, we use all segmentation results as candidate segmentation samples. Then, we arrange 10 observers to screen out the positive and negative samples in the candidate segmentation images, based on criteria in [11]. The positive samples are determined by following standards: (1) the same regions of an image are consistent and uniform; (2) the interior of the region should be simple without many holes; (3) adjacent regions should have significant differences in characteristics while satisfying the regional consistency; (4) the boundary of each region should be simple but not rough, and the spatial position should be accurate. Therefore, and negative samples are those obeying these conditions. Based on the above steps, we take the majority votes to obtain the 1st-rank, the 2nd-rank positive samples and the 1st-rank negative samples for each image in the data set. Finally, we use the 1st-rank pairs to construct the training set (7937 images) and the 1st-rank test set (1058 images) by the random way. For each image, there is one pair of segmentations labeled as positive and negative respectively. We also create the 2nd-rank test set (the same 1058 images) with the 2nd-rank positive and 1st-rank negative pairs corresponding to group a and group b in Fig. 2 respectively. Moreover, the training and test sets were randomly selected from the VOC dataset, which contained 21 categories. Therefore, there is some similarity between them. The evaluation task is more challenging for the 2nd-rank test set, due to the smaller distances between the positive and the negative samples.

IV. EVALUATION

A. Experimental Configuration

For segmentation quality evaluation model, we choose the U-Net as the backbone network. The initial weights of this network are pre-trained parameters on the ImageNet dataset. Before inputting the segmentation image to the evaluation CNN model, it is randomly cropped to 512 × 512 and normalized for each RGB channel. We set Adam optimizer, 0.5 dropout rate, 0.9 momentum, 0.0005 weight decay. Besides, the initial learning rate is 0.001, the batch size is 8. In the training process, it takes 8 hours on a NVIDIA GeForce 2080TI GPU with 12GB memory.

B. Evaluation Results

The proposed evaluation framework is validated on the dataset introduced in III-B. The feature library is obtained by training on the 7937 original images and their corresponding segmentation results.

1) Comparison to unsupervised evaluation measures: The proposed method requires no ground truth segmentation in the test stage. Therefore, we compare it with six well-known unsupervised evaluation measures, i.e., F [15], $F^{'}$ [2], Q [2], Zeb [5], Ecw [6] and E [32]. The six measures are based on describing the inter- or intra-region similarities to evaluate the quality of segmentation. For example F, $F^{'}$ and Q calculate the average squared color errors inside each region. Zeb uses internal uniformity, while Ecw computes the intra-region color error (i.e. the proportion of misclassified pixels). E uses region entropy as the measure of intra-region uniformity. These five indexes are positively correlated with segmentation effect. However, by definition all of these measures do not consider the high-level semantic feature information. To compare the goodness of these measures, the meta-evaluation [22] is used to assess how well each measure can distinguish different qualities of segmentations. However, the data sets used in the experiments (a) and (b) are slightly different from III-B. For each segmentation image we created use different segmentation algorithms [16-18]. Separately, a subjective evalution is performed in which each human evaluator to select the best and worst segmented image from all segmented images. From

TABLE IV
EVALUATION MEAN SCORES AND VARIANCE FOR GROUND TRUTH SEGMENTATIONS.

measures	**mean**	**variance**
$VI \downarrow$	0.294	0.242
$Dice \uparrow$	0.889	0.011
$PRI \uparrow$	0.914	0.062
$SC \uparrow$	0.882	0.076
$IoU \uparrow$	0.878	0.054
$Ours \uparrow$	0.905	0.008

the best and worst segmented image sets selected by each evaluator, we aggregate the best and worst segmented images selected by the seven eavluators into the best set B and the worst set W. Part of the data set is shown in Fig 2. And we perform the experiments to test: (a) if the measures can find the best segmentations for each image in the test set, (b) if they can find the worst segmentations; and (c) if they can correctly identify the positive samples from the negative ones in the test set. The comparison results are shown in Table II. In all the experiments, the proposed method outperforms the other measures. F, $F^{'}$ and Ecw have the low performance which is no better than a random guess. They tend to favor the under-segmentation or over-segmentation of objects. While the proposed method can correctly find over 70% of the best positive and the worst negative samples among the 7 candidate segmentations of each image. Moreover, it achieves 90% accuracy in differentiating the positive segmentations from the negative ones in the 1st-rank test set.

2) Comparison to supervised evaluation measures: Since evaluation without a ground truth is challenging, most literature adopt a supervised way [22] with human labeling. Popular measures include Intersection-over-Union (IoU), Dice [10], Segmentation Covering (SC) [1], Probabilistic Rand Index (PRI) [29], Variation of Information (VI)[30] and so on. Our method does not require the exact ground truth for a test segmentation, however, it is interesting to study whether it can make a comparable performance to the supervised measures. The experimental results are shown in Table III. For the 1st-rank test set, the evaluation task is relatively easier since the quality differences between the positive and the negative pairs are obvious. We can see that the proposed method achieves the accuracy above 90% for both test sets (i.e, 91.793% and 90.148%, respectively), which are better than most of the supervised measures. IoU has the best performance in the 1st-rank test set, but fail to beat the proposed method in the 2nd-rank test set, which is more challenging for evaluation. Our method only requires pre-trained object features from the same class, so it can easily be extended and applied to the evaluation task without human labeled ground truths.

3) Ground truth evaluation: The ability of identifying high quality segmentation is essential for segmentation evaluation. We thus validate the measures by calculating the evaluation results for ground truths. For a good measure, it should consistently assign good scores to these segmentations. The experimental results are shown in Table IV. For supervised measures, ground truths of other images in the same object class are used as the reference for evaluation. We can see that there are five measures producing mean scores close to 1, which is the upper bound of these measures.

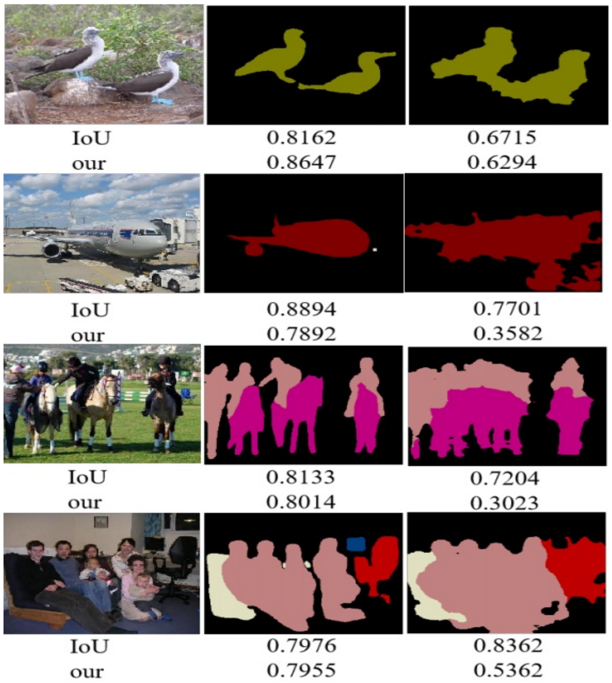

Fig. 3. Images and their segmentations with evaluation scores. From left to right: the original image, the good segmentation and the bad segmentation.

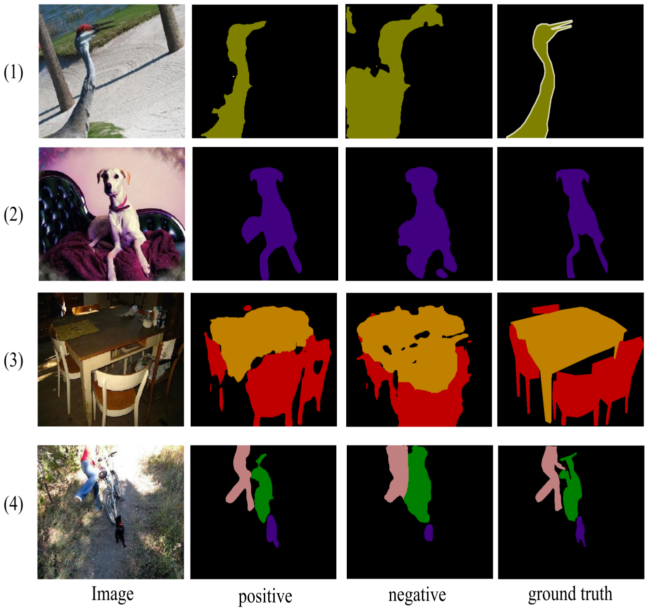

Fig. 4. Segmentation images with one category and multiple categories. From left to right: the original image, the good segmentation, the bad segmentation and the ground truth.

TABLE V
EXPERIMENTAL RESULTS FOR ONE CATEGORY AND MULTIPLE CATEGORIES SEGMENTATION IMAGES.

measures	1		2		3		4	
	positve	negative	positve	negative	positve	negative	positve	negative
$IoU \uparrow$	0.7347	0.3611	0.6498	0.5618	0.7802	0.6760	0.7614	0.1084
$Dice \uparrow$	0.8471	0.5306	0.7877	0.7194	0.8765	0.8067	0.8654	0.7722
$SC \uparrow$	0.9271	0.6451	0.8869	0.8401	0.7293	0.5634	0.9431	0.8949
$PRI \uparrow$	0.9450	0.6937	0.8928	0.8476	0.8168	0.6987	0.9506	0.9117
$VI \downarrow$	0.3896	1.0744	0.5245	0.6522	1.5208	1.9434	0.3809	0.5702
$Our \downarrow$	***0.8092***	***0.5347***	***0.8211***	***0.5243***	**0.5639**	**0.3464**	**0.8688**	**0.4484**

Our measure obtains the second highest mean value and the smallest variance for the ground truths. So it shows the ability of stably evaluating the good segmentations with high scores. A reason for this is that supervised measures mainly calculate the labeling distance between the reference image and the segmentation, while this distance could be large for in-class objects. By representing the objects in the feature space, the variance of in-class objects is reduced, which brings a consistent evaluation result for good segmentations.

4) Qualitative evaluation: To intuitively check the perfomance of our method, we conduct three experiments to verify it. Firstly, we show some qualitative comparisons with the widely used IoU measure in Fig. 3. In the figure, two segmentations in different qualities are shown for each image, as well as their evaluation scores. Ideally, a large difference between the good (the middle) and the bad (the right) segmentations is preferred. In these examples, the proposed measure produces much lower scores for bad segmentations, especially in the 2nd, 3rd and the 4th rows. Compared to IoU which mistakenly assigns higher score to the bad segmentation for the 4th image, our measure correctly identify the under-segmentation result with a lower score.

In image segmentation, the difficulty of segmentation partly depends on the number of objects in the images, and the segmentation quality evaluation is also affected by the category [25]. We employ popular supervised measures [1], [10], [29], [30] to evaluate segmentation image with one category and multiple categories. Experimental images are shown in Fig. 4 and experimental results are shown table V. In the figure, positive and negative samples represent segmented images of different quality and an excellent evaluate measure can be clearly distinguished. However, in the experimental results, other methods do not clearly and accurately distinguish between segmentation images in different quality, especially multiple categories (the third and fourth rows). Besides, We find that other evaluation methods give unreasonable scores for bad segmentation images and often give high scores in multiple categories (i.e. The scores given by Dice, SC, PRI to the third and fourth rows of images). The main reason is that these methods only consider low-level image information. Compared with the proposed method, this method can not only distinguish the segmentation quality of single category or multiple categories images, but also objectively give a feedback to the segmentation image. Meanwhile, this method can evaluate the segmentation quality without being limited by the difficulty of segmentation.

Through experimental analysis, it is concluded that the proposed method is not effective in evaluating the over-segmentation images. From the partial over-segmentation image shown in Fig. 5, it can be seen that the evaluation score given by the proposed method is less objective. If we draw the segmentation effect of the image directly from the score, it will cause misjudgment. For example, b in the third row would be mistaken for a better segmentation.

V. CONCLUSION

In this paper, we studied the segmentation evaluation by designing a CNN based evaluation framework. A feature embedding module which integrates meta evaluation principle was proposed and trained in a metric learning process. In the test stage, we used the trained object features to calculate the cosine distance and obtain the score for the segmentation result. Also, an object segmentation evaluation dataset was constructed to validate the proposed method. We compared the proposed method with different unsupervised and supervised measures, which showed that it can provide more accurate evaluation results for segmentations of different qualities on the proposed evaluation dataset.

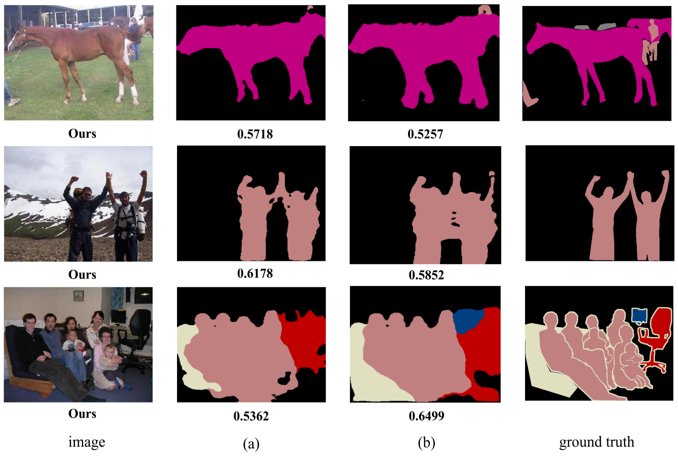

Fig. 5. Badly evaluated images and their segmentation with evaluation scores. From left to right: the original image, two segmentation images and ground truth.

VI. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Nos. 61772435, 61961038), Sichuan highway science and technology project (No. 2019-01) and the Fundamental Research Funds for the Central Universities (No. 2682021ZTPY069).

REFERENCES

[1] P. Arbelaez, M. Maire, C. C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(5):898–916, 2011.

[2] M. Borsotti, P. Campadelli, and R. Schettini. Quantitative evaluation of color image segmentation results. Pattern Recognition Letters, 19(8):741–747, 1998.

[3] Z. Cai, Y. Liang, and H. Huang. Unsupervised segmentation evaluation: an edge-based method. Multimed Tools and Applications, 76:11097C11110, 2017.

[4] A. Cavallaro, E. Gelasca, and T. Ebrahimi. Objective evaluation of segmentation quality using spatio-temporal context. In International Conference on Image Processing, page 49C56, 2002.

[5] S. Chabrier, B. Emile, H. Laurent, C. Rosenberger, and P. Marche. Unsupervised evaluation of image segmentation application to multispectral images. In Proceedings of the 17th International Conference on Pattern Recognitionn, pages 576–579, 2002.

[6] H. Chen and S. Wang. The use of visible color difference in the quanitative evaluation of color image segmentations. In IEEE international conference on acoustics, speech, and signal processing, pages iii–593, 2004.

[7] L. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking Atrous Convolution for Semantic Image Segmentation. Technical report, Google LA, Dec 2017.

[8] Y. Chen, D. Dai, J. Pont-Tuset, and L. Gool. Scale-aware alignment of hierarchical image segmentation. In IEEE Conference on Computer Vision and Pattern Recognition, pages 364–372, 2016.

[9] H. Christensen and P. Phillips. Empirical evaluation methods in computer vision. World Scientific Publishing Company, 2002.

[10] L. Dice. Measures of the amount of ecologic association between species. Ecology, 26(3):297–302, 1945.

[11] R. Haralick and L. Shapiro. Image segmentation techniques. Computer vision, graphics, and image processing, 29(1):100–132, 1985.

[12] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask r-cnn. In IEEE International Conference on Computer Vision, pages 2961–2969, 2017.

[13] C. Huang, Q. W. Q, and F. Meng. Qualitynet: Segmentation quality evaluation with deep convolutional networks. In IEEE Visual Commnunications and Image Processing, pages 1–4, 2016.

[14] J. Lahoud, B. Ghanem, M. R. Oswald, and M. Pollefeys. 3d instance segmentation via multi-task metric learning. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 9255–9265, 2019.

[15] J. Liu and Y. Yang. Multi-resolution color image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 16(7):689C700, 1994.

[16] Y. Liu, P. Jiang, V. Petrosyan, S. Li, J. Bian, L. Zhang, and M. Cheng. Del: Deep embedding learning for efficient image segmentation. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, pages 864–870, 2018.

[17] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In IEEE Conference on Computer Vision and Pattern Recognition, pages 3431–3440, 2015.

[18] M. Meila. Comparing clusterings: an axiomatic view. In International Conference on Machine Learning, pages 577–584, 2005.

[19] V. Movahedi and J. H. Elder. Design and perceptual validation of performance measures for salient object segmentation. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, page 49C56, 2010.

[20] B. Peng, M. Simfukwe, and T. Li. Region based image segmentation evaluation via perceptual pooling strategies. Machine Vision and Applications, 28:477–488, 2018.

[21] J. Pont-Tuset and F. Marques. Measures and meta-measures for the supervised evaluation of image segmentation. In IEEE Conference on Computer Vision and Pattern Recognition, pages 2131–2138, 2013.

[22] J. Pont-Tuset and F. Marques. Supervised evaluation of image segmentation and object proposal techniques. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(7):1465–1478, 2016.

[23] X. Ren and J. Malik. Learning a classification model for segmentation. In IEEE International Conference on Computer Vision, pages 10–17, 2003.

[24] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241, 2015.

[25] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In IEEE Conference on Computer Vision and Pattern Recognition, pages 815–823, 2015.

[26] R. Shi, K. N. Ngan, S. Li, P. Raveendran, and H. Li. Visual quality evaluation of image object segmentation: Subjective assessment and objective measure. IEEE Transactions on Image Process, 24(12):5033–C5045, 2015.

[27] R. Shi, K. N. Ngan, S. Li, P. Raveendran, and H. Li. Visual quality evaluation of image object segmentation: Subjective assessment and objective measure. IEEE Transactions on Image Process, 24(12):5033C5045, 2015.

[28] W. Shi, F. Meng, and Q. Wu. Segmentation quality evaluation based on multi-scale convolutional neural networks. In IEEE Visual Communications and Image Processing, pages 1–4, 2017.

[29] R. Unnikrishnan, C. Pantofaru, and M. Hebert. A measure for objective evaluation of image segmentation algorithms. In Workshop on Empirical Evaluation Methods in Computer Vision, IEEE Conference on Computer Vision and Pattern Recognition, page 34, 2005.

[30] P. Viola and W. Wells. Alignment by maximization of mutual information. International Journal of Computer Vision, 24(2):137–C154, 1997.

[31] S. Wang. New benchmark for image segmentation evaluation. Journal of Electronic Imaging, 16(3):033011, 2008.

[32] H. Zhang, J. Fritts, and S. Goldman. An entropy-based objective segmentation evaluation method for image segmentation. In SPIE Sotrage and Retrieval Methods and Applicaitons for Multimedia, pages 38–49, 2004.

[33] Kulis B. Metric learning: A survey[J]. Foundations and trends in machine learning, 2012, 5(4): 287-364.

[34] Yang L, Jin R. Distance metric learning: A comprehensive survey[J]. Michigan State Universiy, 2006, 2(2): 4.