Improved Similarity Propagation for One-Shot Semantic Segmentation

Siddhartha Gairola*[1], Ayush Chopra[2], Mayur Hemani[2], and Balaji Krishnamurthy[2]

[1]IIIT Hyderabad
[2]Media and Data Science Research, Adobe Experience Cloud

Abstract

Deep neural-network methods for one-shot segmentation have improved in accuracy but are significantly worse than supervised methods. This work demonstrates that inadequate similarity propagation in existing one-shot segmentation methods is a major reason for this. We propose two strategies for improving the sharing of similarity information between the support and query inputs - a preconditioning step using a transformed variant of the support image as a query, and a cyclic reinforcement training constraint. We also introduce a more nuanced reporting framework, with an evaluation dataset H-PASCAL-5^i, for improved clarity in evaluating the performance of one-shot segmentation methods. The proposed method demonstrates state-of-the-art performance on both PASCAL-5^i and H-PASCAL-5^i for the one-shot setting.

1. Introduction

Semantic image segmentation assigns class labels to pixels in an image. It finds applications in image editing [3, 2, 1, 26, 29, 30], medical diagnosis [21, 19, 10, 27], automated driving [9] etc. Supervised deep neural network methods such as [32, 4, 5, 6, 7, 8, 38] allow for highly accurate image segmentation for a small number of fixed classes of objects that are used for training the methods. Class-agnostic segmentation extends the semantic image segmentation task for object classes not used for training. They predict a binary segmentation mask with ones at pixels of objects of a target class and zeroes elsewhere. One-shot (or few-shot) image segmentation methods, like OSLSM [24], CANet [36], PANet [33], represent a class of approaches for class-agnostic segmentation that uses a support image and its segmentation mask as additional input to predict the segmentation mask of a given image. In this paper, we present improvements for training existing one-shot segmentation networks by improving the propagation of similarity in them. We also present a more nuanced protocol for reporting performance for these methods. One-shot segmentation (OSS) methods typically have three inputs - a *query* image, a *support* image and a *support mask* (the segmentation mask corresponding to the support image) from which they learn to predict the segmentation mask for the query image. These methods typically leverage the similarity between the query and the support images to selectively decode features from pre-trained networks to output the segmentation mask Figure 1. However, the propagation of the similarity information across the networks used is not accurate as evident from the low mean intersection-over-union (mIoU) of OSS methods [36, 33].

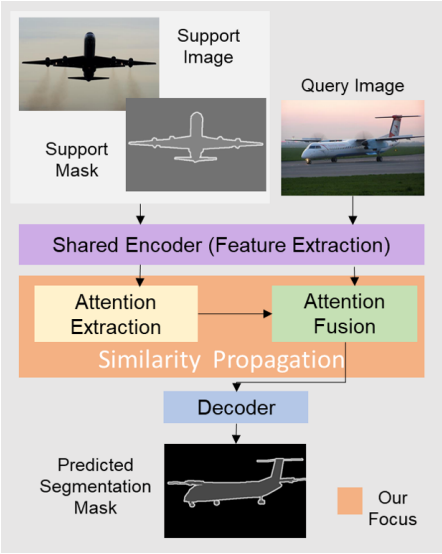

Figure 1: Typical architecture of One-shot Image Segmentation networks. Similarity propagation enables the network to identify the target regions in the query image for segmentation. We demonstrate improved flow of similarity information and as a result improved one-shot segmentation.

Most current work on one-shot image segmentation focuses on producing novel neural network architectures (see Section 2). We believe that existing network architectures

*work done as part of Adobe MDSR internship program

can be refined to produce better output by identifying their shortcomings and rectifying them. We posit that regions misidentified by existing OSS methods consist of pixels that are: (a) not captured by the similarity propagation scheme by these methods, or (b) not considered by the class-conditional similarity matching semantic at all, or (c) in regions that are inherently hard to segment and missed even by supervised methods like [8]. The focus of this work is to establish the presence of the similarity propagation gap and bridging it by improving the constraints that enforce similarity propagation, and by introducing a network conditioning step prior to the actual training.

Our experiments with [36] and [25] reveal gaps in similarity propagation (see Section 3.1). We measure the similarity of regions of errors by [36] across image pair samples of identical classes and determine it to be high. The segmentation output for identical support and query inputs is also of low quality. These results indicate that the class and visual similarity information is not propagated optimally across the images. Towards rectifying this issue, we condition the network initially with the support and query pairs where the query image is a transformed version of the support image. Further, we add a cyclic reinforcement constraint on the network with two stages at every iteration - the first stage consumes a support image, a support mask, and a query image and predicts the query mask; the second stage flips the inputs (support image is the new query image), and uses the predicted query mask from the first stage as the support mask. This additional requirement reinforces the sharing of similarity information by the network.

With these two improvements - the transformed-pair conditioning and the cyclic reinforcement constraint in end-to-end training - we achieve state-of-the-art results. Details of the experiments establishing the premise as well as ablation for the individual and combined improvements are included in the paper.

The proposed method is evaluated using the mean intersection-over-union (mIoU) of the output segmentation mask with the ground-truth mask. The results are computed on partitions of the PASCAL 5^i dataset [24] (named H-PASCAL 5^i). We observe that random sets of support and query pairs are used for quality measurement of one-shot segmentation leading to discrepancies in reporting performance. The output quality also varies with the inherent difficulty of the segmentation task even with supervision (as detailed in Section 3.1). Accordingly, we partition the evaluation sets based on the visual similarity between the query and support image pairs, on the performance of a state-of-the-art supervised method [8], and on the extent of the object to be extracted. This ensures more nuanced reporting and directs progress towards the pragmatic goal of real-world one-shot segmentation. We also release the evaluation splits used for computing the reported metrics. The paper also includes detailed ablation studies and comparisons with baseline one-shot segmentation methods.

To summarize, the paper makes the following contributions:

1. Analysis of the gaps in similarity propagation in one-shot segmentation methods.
2. A pre-training conditioning of networks with transformed support images as query to improve output on highly similar images.
3. A cyclic reinforcement constraint by cascaded end-to-end training with flipped inputs, to improve the propagation of similarity information in an existing one-shot segmentation network.
4. A nuanced evaluation framework for improved performance reporting, and the corresponding evaluation dataset partitions H-PASCAL 5^i (to be released after the conference).

The contributions presented in the paper are complementary to the efforts made towards designing novel network architectures, are applicable to many contemporary networks for the task of one-shot segmentation, and outperform the state-of-the-art.

The next section discusses previous work related to the problem and the solutions proposed. Section 3 presents evidence for the need for the proposed improvements as well as their details, and introduces the new evaluation framework. Section 5 presents results on both the prevailing measurement framework as well as the proposed hierarchical partitioned evaluation framework. Section 6 presents deeper analysis to corroborate the claim made by the proposed improvements for similarity propagation, and ablation studies.

2. Related Work

In recent years, one-shot image segmentation has become an active area of research in computer vision. In this section, we summarize some of the key works in this area, highlight their shortcomings and place our work in context of the state of the art. We make a distinction between one-shot and few-shot segmentation for the scope of this paper. Methods for few-shot segmentation are discussed (and evaluated) only in the context of their one-shot performance.

One-shot Semantic Segmentation Shaban et al. [24] introduce a dual-branched neural network model for one-shot image segmentation with a support and query branch. The support branch is conditioned on the support input to predict the weights of the last layer in the query branch which then predicts the query mask. Rakelly et al. [20] improve upon [24] by employing a late fusion strategy for the support mask and support feature maps to segment the query image. These methods typically adopt a parametric module in a two branch setting, which fuses information extracted from the support set to produce the segmentation mask. The two-branches used have different structures which makes the training and evaluation procedure cumbersome.

The idea of prototypical networks [28] is adopted by [11], to approach few-shot segmentation using metric learning. However, the model is complex with a three stage

training procedure involving complicated training configurations. Additionally, their method uses the prototypes as guidance to fix the query set segmentation rather than obtaining segmentation directly from metric learning. Zhang et al. [37] introduce masked average pooling (MAP), and estimate similarity between the pooled support features to estimate similarity with the query features for predicting the query's segmentation maps.

Analogous to [11], Wang et al. [33] draw inspiration from prototypical learning [28] and adopt late fusion [20] and MAP [37] to incorporate the support masks with updating annotations to segment images. Recent methods on few-shot segmentation build upon [37] to use MAP more effectively, [36] adopt a learnable method through an attention mechanism to fuse information from multiple support examples along with iterative refinement, on the other hand [18] employs weighing of background/foreground features and boosting inference with an ensemble of guides.

These methods for one-shot segmentation depend on the support set to produce the segmentation for the query image, but fail to utilize the support to its fullest as discussed in Section 3.1. The work presented in this paper proposes to bridge this gap.

Cycle Consistency Cycle consistency is the idea of using transitivity as a means to regularize ill-posed objectives in optimization setups. It has been used for dense semantic alignment [40, 39], depth estimation [13], 3D shape matching [17], co-segmentation [31] and structure from motion [35]. Most recently, cycle consistency was used for unpaired image-to-image translation in [34, 41], where a *cycle consistency loss* is introduced to train a GAN [14]. That work is inspired by dual learning used in machine translation [15]. Inspired by cycle consistency, we propose a cyclic reinforcement constraint (CRC) to augment the training of one-shot segmentation networks. The proposed CRC is formulated specifically for the OSS setting and introduce in Section 3. In the present work, we demonstrate that the addition of this constraint improves the sharing of similarity information between the support and query images (Section 5 and Section 6).

3. Method

In this section, we first establish firm ground to motivate our work, and then elucidate the proposed method in the one-shot segmentation setting.

3.1. Premise Validation

The improvements we propose are based on the premise that there are gaps in similarity propagation. We first present experimental evidence validating this assumption. The experiments use author-provided implementations of the most recent state-of-the-art methods [36] and [25].

Errors by OSS methods versus Supervised Segmentation methods Some image regions are inherently hard to segment and missed even with strong supervision. We posit that OSS methods make errors for regions besides these hard regions. Table 1 presents the overlap between the error regions of output masks from two OSS methods ([36] and [25]) with the error regions of masks produced by DeepLab v3+ [8] (a state-of-the-art supervised method). The overlap between the error regions for OSS methods and DeepLab v3+ is very small, while the gap in the mean-IoU for predictions (DeepLab v3+ versus OSS methods) is large. This indicates that DeepLab v3+ can overcome a substantial fraction of errors made by the OSS methods. Therefore, those regions of errors are not characteristically difficult to segment.

OSS Method	FN Overlap with DLv3+ (mIoU)	FP Overlap with DLv3+ (mIoU)	Prediction mIoU Gap
CANet	18.32%	9.11%	23.28%
AMP	10.99%	19.06%	39.41%

Table 1: Mean intersection-over-union measured over erroneous predictions by CANet [36] and Adaptive Masked Proxies [25] measured against those from DeepLab v3+ [8]. False Negative (FN) regions are the regions missed by the segmenter, while False Positive (FP) regions are the ones that they mis-predict as part of the mask. These numbers are obtained from 400 test image pairs from the PASCAL VOC 2012 dataset ([12]).

Similarity of mis-predicted regions for different OSS methods We determine the similarity between regions mis-predicted by [36] through a second experiment. We compute the Masked Average Pooling (MAP) vectors (as described in [37]) for the following two regions of each image for several pairs of images of identical classes:

a. the pixels covered by the ground-truth masks (MAP_{gtmask}), and
b. the pixels present in the regions mis-predicted by the OSS method (MAP_{error}).

For each image pair (A, B) of a class, the ratio of cosine similarity between the MAP vectors for the error regions ($MAP_{error}(A).MAP_{error}(B)$) and that between the MAP vectors of the ground-truth mask regions ($MAP_{gtmask}(A).MAP_{gtmask}(B)$) is a measure of the similarity of the corresponding error regions relative to the similarity of the ground-truth mask regions. We measure the ratio over a 1000 pair of images and find the average to be 0.87 (standard deviation = 0.25). The high value of relative similarity of error regions substantiates the claim that errors are committed in regions of similarity that could have possibly been propagated from the support to the query.

Identical Inputs: (Support = Query) Table 2 reports results from a third experiment with ([36] and [25]) where the same image is given as input for both support and query

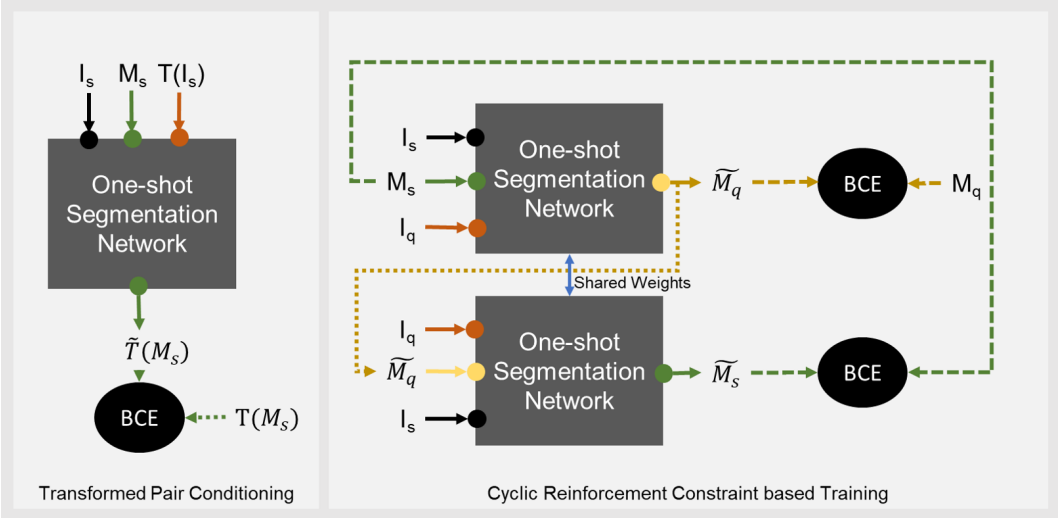

Figure 2: Transformed-Pair Conditioning and Cyclic Reinforcement Constraint based training explained schematically. I_s, M_s, I_q, M_q are the support image, support mask, query image and query mask respectively. $T(I_s)$ is the transformed support image and $\tilde{T}(M_s)$ is the corresponding mask. BCE is the binary cross-entropy loss.

Figure 3: Examples of transformed image pairs used in the pre-conditioning step. Details of the transformations are provided in the supplementary material. **Top Row:** Distortion and Pepper Noise. **Bottom Row:** Rotation and Zooming-in.

(including the ground-truth mask for the support). This constitutes a the most basic test for similarity propagation in the network. Ideally, the network must produce the exact mask as provided in the support input, because the query and the support are as similar as possible. However, both methods ([36] and [25]) perform poorly for these inputs indicating the loss of similarity information in the networks.

OSS Method	Fold 0	Fold 1	Fold 2	Fold 3
CANet	43.22%	49.50%	40.21%	41.73%
AMP	49.06%	68.55%	67.45%	59.36%

Table 2: Mean-IoU measured for OSS methods with query image = support image for test images from PASCAL 5^i dataset.

Considered together, the results of these experiments indicate that similarity propagation is not adequately exploited by existing OSS methods.

3.2. Conditioning with Transformed-Pair Images

We submit that one reason for the poor performance of OSS methods on identical query and support images is that they focus exclusively on class-conditional similarity. Most OSS methods employ backbone networks pre-trained on classification tasks (e.g. [36] employs ResNet-50 ([16] pre-trained on ILSVRC [22] dataset). These networks are trained to produce feature representations necessary to identify the class of an object. For two images with segmentation targets of identical class, this information may be similar in specific regions of the object but not necessarily entire objects ([23] provide a means to visualize this). To alleviate this shortcoming, we pre-train the layers of the network subsequent to the feature-producing layers with near-identical query and support pairs. The query image is augmented with various noises (Gaussian and pepper-noise) and transformations (flips, rotations, and non-linear distorts). Figure 3 shows examples of such distorted training pairs.

The result of this conditioning of the network layers prior

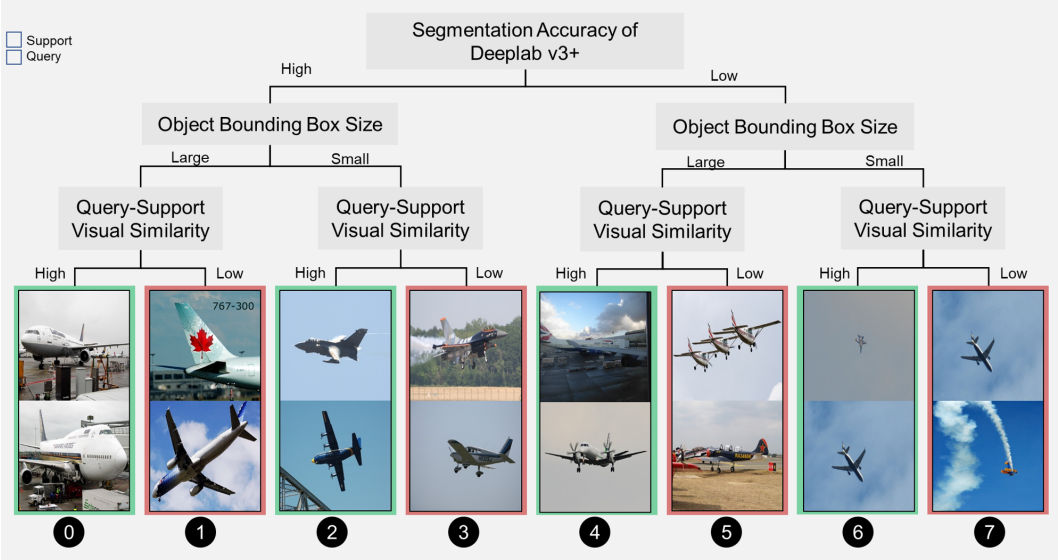

Figure 4: Example of images from the H-PASCAL 5^i dataset. The query is chosen randomly from each partition, while the support is chosen by a visual similarity search over the corresponding bounding-box size partition (k-nearest and k-furthest). For the L0 set of pairs, for example, the query images are relatively easy to segment using DeepLabv3+ and have relatively large objects, and the support image is visually very similar to the query.

to the training is that the actual training converges faster (See Figure 5).

3.3. Cyclic Reinforcement Constraint

One-shot segmentation methods typically predict only the query mask given the support image and mask, and the query image. Methods like [36] also apply iterative correction over their output to further improve the quality. One possibility for improved similarity propagation is to also predict the support mask from the same network. Adding this symmetric task would ensure that the similarity between the support and query images is thoroughly utilized by the network. However, given the structure of the network it is easy for the branch predicting the support mask to trivially collect it from the support input.

In order to alleviate this issue, and taking a cue from the cycle-consistency loss introduced in [41] we recognize that adding a transitivity loss enables better regularization. However, unlike in [41], our goal is to predict only a part of the input - the support mask, as a secondary output to squeeze out the class-conditional similarity information between the support and the query images. To this end, the network from [36] is arranged in a cascaded fashion (Figure 2) where the first stage predicts the mask for the query image, which is then used as the support mask in the second stage to predict the original support mask. The network is trained end-to-end with the binary cross-entropy loss over the output of each stage. The cyclic constraint is formulated as:

$$\begin{aligned} \tilde{M}_q &= F(I_s, M_s, I_q) \\ \tilde{M}_s &= F(I_q, \tilde{M}_q, I_s) \end{aligned} \tag{1}$$

$$\begin{aligned} Loss_q &= BCE(M_q, \tilde{M}_q) \\ Loss_s &= BCE(M_s, \tilde{M}_s) \\ Loss_{total} &= Loss_q + Loss_s \end{aligned} \tag{2}$$

Here I_s, I_q, M_s and M_q are the support image, query image, support mask and the ground-truth query mask respectively. $\tilde{M}_q$ and $\tilde{M}_s$ are the predicted query and support masks respectively. And BCE is the binary cross-entropy loss function over the pixels of the produced masks. The trained network is tested with the different PASCAL 5^i dataset [24] splits.

4. Experimental Settings

4.1. Dataset

We report our results for the test classes of the PASCAL 5^i dataset [24]. Evaluation of [36] and [25] (code made available by authors) reveals that one-shot segmentation methods work better when the target objects are visually similar besides belonging to the same class (see Table 3). In practical applications too, a user may be expected to provide as support, the segmentation annotation for an image that is similar to the query. The current evaluation schemes as described in [24] and later revised in [20] do not consider these nuances of visual similarity. We propose to augment the protocol from [24] with a partitioning scheme for the test data and name it H-PASCAL 5^i. We partition the test pairs of the PASCAL 5^i dataset into eight groups based on three dimensions:

1. The accuracy with which a pre-trained supervised semantic segmentation network (DeepLab v3+ [8]) pro-

duces the segmentation mask of the query image.
2. The extent of the target objects relative to the query image size.
3. The visual similarity between the query and the support images.

We use the L_2 distance between features from a pre-trained ResNet-101 for determining visual similarity. For each test class in a fold, the mean intersection-over-union is computed for the masks generated by the DeepLab v3+ network for each image. Images with mIoU one standard deviation higher and lower than the mean are classified as easy and hard to segment respectively. Each partition is further subdivided similarly by the relative size of the target object (area of the bounding box). Query images are randomly chosen from each partition, and top k closest and furthest images from the corresponding partition are chosen as support. Figure 4 illustrates the constitution of the partitions. The detailed statistics of the new query-support protocol is included in the supplementary material.

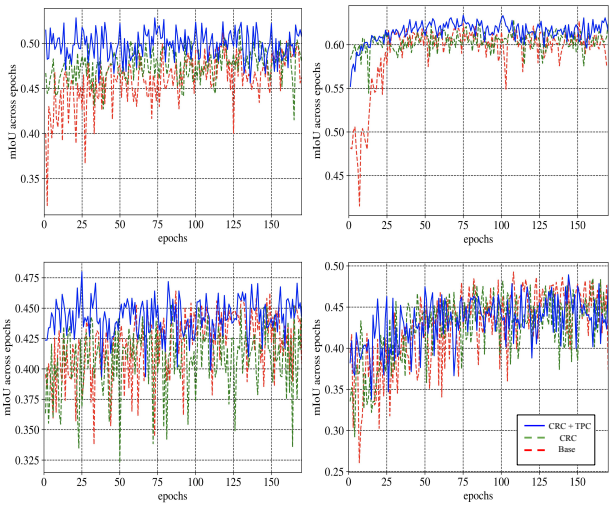

Figure 5: Test mIoU evolution over epochs for the different PASCAL-5i splits in training with transformed-pair conditioning and cyclic-reinforcement constraint. From Top-Left (clockwise): split-1, split-2, split-4, split-3. (Zoom in to see the plots clearly)

4.2. Metric and Reporting Protocol

We follow the protocol established by Shaban et al. [24]. The mean intersection-over-union metric for output masks with respect to the ground-truth mask is reported. Different instances of the network are trained with the different training splits of the PASCAL 5^i dataset, each with 15 classes of objects and their masks. Testing is done on the images with objects of the 5 withheld classes. We report the mean IoU values for random support and query pairs (images of the same class), as well as for pairs from H-PASCAL 5^i as described in the previous section.

4.3. Implementation Details

The proposed framework is generic and can be integrated with any one-shot segmentation network architecture. For our experiments in this paper, we employ the recent state-of-the-art method [36]. The experiments use the author-provided implementation of the CANet ([36]) network. The transformed-pair conditioning step is performed for 30 epochs before training for 200 epochs with the distinct image pairs from the PASCAL 5^i dataset [24]. The training is done on virtual machines on Amazon AWS with four Tesla V100 GPUs and 16-core Xeon E5 processors. The learning rate is kept at a low value (2.5×10^{-5}) and the batch size is 24 for all training. To validate generality of the approach, we provide comprehensive evaluations with other recent (top-performing) one-shot segmentation architectures [33, 25] in the supplementary material.

5. Results

We report our results using the hierarchical partitions from the proposed H-PASCAL 5^i dataset. (defined in Section 4.1). The relative performance on the different partitions provides insights into the behavior of the one-shot segmentation networks vis-à-vis the characteristics of the input images. To provide a comprehensive analysis, we (here) report comparisons against the most recent state-of-the-art method- CANet ([36]). Comparisons with other recent methods ([33], [25]) are included in the supplementary material. For completeness, we also compare our performance using the prevailing method (from [24]) that employs random query and support image pairs (partitions) from each test class in PASCAL 5^i splits. In both cases, the proposed framework outperforms the current state-of-the-art method.

Evaluation with Hierarchical Partitions of H-PASCAL-5^i The proposed hierarchical partitions have 8 levels of tasks with differing ease of segmentation, object sizes and visual similarity between the support and query images. L0, L1, L2, L3 represent an easy set and L4, L5, L6, L7 represent a hard set for the segmentation task. The even partitions (L0, L2, L4, L6) represent images with relatively high visual similarity within their parent partitions. L0, L1, L4, and L5 have relatively larger objects compared to their counterparts (L2, L3, L6, L7). This hierarchy is outlined in Figure 4 and explained in Section 4.1.

These distinctions between the tasks enables the correlation of one-shot segmentation performance to characteristics of the images. For example, the average of the mean IoU values across splits for the CANet [36] (Base) listed in Table 3 indicate that the network has consistently higher performance for larger objects (L0,L1 in the easy set, L4, L5 in the hard set) than for smaller objects (L2, L3, L6, and L7).

The effect of the combined transformed-pair conditioning and cyclic reinforcement constraints is evident from the

	split-1		split-2		split-3		split-4		mean	
Partition	Base	CRC + TPC	Base	CRC + TPC	Base	CRC + TPC	Base	CRC + TPC	Base	CRC + TPC
L0	61.64	**64.59**	77.24	**77.57**	51.40	**52.45**	50.27	**53.60**	60.14	**62.05**
L1	62.38	**64.50**	71.86	**72.76**	**55.51**	55.22	**61.83**	60.17	62.89	**63.16**
L2	59.43	**60.81**	57.56	**58.85**	49.18	**49.58**	44.79	**45.05**	52.74	**53.57**
L3	61.41	**63.15**	60.38	**61.03**	46.97	**48.44**	**50.85**	50.77	54.90	**55.85**
L4	26.27	**27.45**	39.11	**41.10**	30.80	**33.11**	**32.22**	30.99	32.10	**33.16**
L5	24.88	**25.70**	38.54	**38.87**	31.22	**32.79**	30.08	**30.87**	31.18	**32.06**
L6	**17.99**	13.89	17.51	**21.38**	18.44	**24.25**	14.45	**16.13**	17.10	**18.91**
L7	8.46	**8.95**	19.72	**23.43**	24.69	**25.17**	15.84	**20.85**	17.18	**19.60**
CMPS	40.29	**41.13**	47.74	**49.38**	38.52	**40.26**	37.54	**38.55**	41.02	**42.38**

Table 3: One-shot results under the meanIoU evaluation metric computed on the proposed hierarchical evaluation set. Shows superior performance of using the proposed CRC+TPC training scheme across splits(1-4) and partitions(L0-L7) over CANet (Base) [36]. A **CMPS** (Combined Mean Partition Score) is the mean value computed across partitions for each split. The best results are highlighted in bold.

Method	split-1	split-2	split-3	split-4	mean
CRC + TPC (ours)	**52.80**	**63.39**	**48.32**	48.97	**53.37**
CA-Net [36]	51.17	61.40	46.40	**49.25**	52.05
PA-Net [33]	42.40	58.00	51.10	41.20	48.13
SG-One [37]	40.20	58.40	48.40	38.40	46.30
co-FCN [20]	36.70	50.60	44.90	32.40	41.10
OSLSM [24]	33.60	55.30	40.90	33.50	40.80
AMP [25]	35.20	45.70	39.30	34.20	38.60

Table 4: Mean IoU of one-shot segmentation on the PASCAL 5^i dataset using random partitions. The best results are highlighted in bold. Using CRC + TPC with Base ([36]) achieves state-of-the-art performance.

Method	split-1	split-2	split-3	split-4	mean
Base	40.29	47.74	38.52	37.54	41.02
CRC	40.89	49.08	39.60	37.88	41.87
CRC + TPC	**41.13**	**49.38**	**40.26**	**38.55**	**42.38**

Table 5: Effect of adding CRC (only) and CRC with TPC on performance (mIoU) computed on the hierarchical evaluation set (averaged across splits(1-4) and partitions(L0-L7)).

	Support = Query				
Method	split-1	split-2	split-3	split-4	mean
Base	43.22	49.50	40.21	41.73	43.66
CRC+TPC.	**44.13**	**53.10**	**42.35**	**45.71**	**46.33**

Table 6: Mean IoU for vanilla CANet (Base) and CANet + Proposed improvements with identical support and query images. The results indicate that the proposed improvements (transformed-pair conditioning (TPC), and cyclic-reinforcement constraint (CRC)) result in improved similarity propagation.

high mean intersection-over-union values listed in Table 3 across splits(1-4) and partitions(L0-L7). Overall, our approach gets an average gain of about 1.36% over the base network across all splits and partitions. The gain is significantly higher (about 1.92%) for L0 which represents easy to segment query images with large objects and high similarity between the query and support images. Also for the easy set (L0-L3), the proposed method consistently outperforms the base CANet network on average mIoU across levels. A surprising discovery is the consistent gain in the L7 partition, and it indicates that besides improving the similarity propagation, the proposed approach also improves performance on the most difficult cases with sparse query support similarity. Further, results in Table 5 highlight that using CRC and TPC individually also results in a performance gain. For instance, on using CRC only, mIoU over all splits improved from 41.02% to 41.87% and this further improved to 42.38% when TPC was also used.

Evaluation with Random Partitions of PASCAL 5^i The proposed framework result in an average mean intersection-over-union of 53.37% for the PASCAL 5^i dataset splits, a gain of 1.32% over the state-of-the-art. Table 4 compares the performance of our method over the reported and measured for existing one-shot segmentation methods [1].

[1] Some values in Table 4 may differ from ones reported in original reference paper even though we used author-provided implementations (or checkpoints) to report performance.

	split-1		split-2		split-3		split-4		mean	
Partition	Base	CRC + TPC	Base	CRC + TPC	Base	CRC + TPC	Base	CRC + TPC	Base	CRC + TPC
L0	45.40	**61.31**	29.99	**34.10**	20.97	**21.85**	**42.24**	40.33	34.65	**39.40**
L1	47.50	**61.40**	**36.42**	35.15	**30.50**	30.27	48.83	**50.96**	40.81	**44.44**
L2	40.19	**45.03**	12.92	**17.56**	17.60	**17.95**	**34.78**	33.84	26.37	**28.59**
L3	49.42	**55.95**	19.21	**21.85**	14.44	**15.62**	37.01	**40.19**	30.02	**33.40**
L4	14.68	**22.21**	11.51	**16.80**	12.63	**13.34**	19.37	**22.45**	14.55	**18.70**
L5	16.55	**21.61**	9.80	**15.01**	20.22	**20.58**	**22.42**	20.02	17.24	**19.31**
L6	**14.00**	11.55	7.80	**09.16**	4.35	**7.75**	**14.43**	13.27	10.14	**10.43**
L7	**6.98**	5.11	8.94	**13.29**	9.95	**10.08**	**13.89**	13.69	9.94	**10.54**
CMPS	29.34	**35.52**	17.07	**20.37**	16.33	**17.18**	29.12	**29.34**	22.96	**25.60**

Table 7: Mean IoU for vanilla CANet [36] (Base) and CANet + Proposed improvements (CRC+TPC) with the support mask set to 0. The results indicate that the proposed improvements (transformed-pair conditioning (TPC), and cyclic-reinforcement constraint (CRC)) result in improved similarity propagation. Combined Mean Partition Score (**CMPS**) is the mean value computed across partitions for each split. Mean CMPS improved from 22.96 to 25.60 on using TPC and CRC

6. Analysis

To further probe the capabilities of our contributions, we compare the behavior of the vanilla CA-Net [36] (Base) network with the one trained with the proposed improvements (CRC and TPC). Specifically, we determine the gain in the utilization of the similarity information from the support image by two measurements.

6.1. Measuring Similarity Propagation

Next, we stress test similarity propogation in two extreme cases a) Where Support Image (and Mask) is same as Query b) When Support Mask is not used (set to 0).

Performance on Identical Support and Query Images The performance of a one-shot segmentation method on identical inputs for the query and the support is about testing the lower bound on the similarity propagation in the network. Table 6 presents the gain in mean intersection-over-union for identical images over the vanilla CANet network. The average gain of 2.65% over all splits indicates clearly that the network is better utilizing the similarity information between the query and support images.

Performance when Support Mask is Not Used The class-conditional similarity information shared between the query and support images comes from the actual pixel-map (RGB) images input to the networks. By removing the support mask altogether, we can assess if the network is indeed learning to propagate the similarity between these images. Table 7 lists the results of removing the support mask at the test time for both the vanilla CANet network, as well as one trained with the proposed improvements. The gain in the mean IoU for our methods indicates that the propensity of the network to propagate similarity has improved. For instance, notice that not supplying the support mask causes a drop of more than 16% in the L0 set for split-1 (from 61.64% to 45.40%) for the base network but only about 3% for our framework. On an average the gain over the base network without a support mask is of about 3.23% (almost 6.25% in split-1). These numbers clearly indicate improvement in similarity propagation between the support and query pairs.

6.2. Effect of Transformed-pair Conditioning

To assess the effect of the pre-training step of conditioning the network by training with (image, transformed-image) pairs, we perform the following ablation studies. The results are presented only for Split-1 test images of the PASCAL 5^i for brevity.

Data augmentation with Transformed Pairs To check if the transformed pairs could simply be mixed with the distinct image pairs at training time, we augment the training data with transformed-pairs and train the CANet network. The resulting mean IoU over test images in Split-1 is 50.74% which is smaller than that of the base network (52.8%) on the same split.

Post-training Fine-tuning It is also possible to fine-tune the trained network with transformed-pair images. To check if that is an effective strategy, we fine-tune the CANet network for 10 epochs after training them on Fold-0. The results have worse mean IoU values (47.04%) indicating that fine-tuning with transformed pairs is not a viable option.

7. Conclusions and Future Work

The paper presents a rigorous argument that similarity propagation in one-shot image segmentation networks is sub-optimal. It proposes two strategies for improving similarity propagation in existing one-shot image segmentation networks. The combined effect of the strategies increases the performance of the network to beyond the current state-of-the-art, shown by a comprehensive set of experiments. The paper also presents a nuanced, unambiguous evaluation dataset for a clearer means of measuring the performance of

one-shot image segmentation networks. Intuitively, class-conditional similarity matching can only determine pixels with a matching class-mix between the query image and the masked region of the support image. In future work, we focus on exploring an "objectness" prior (as an objective) during training.

References

[1] Yağiz Aksoy, Tunç Ozan Aydin, Marc Pollefeys, and Aljoša Smolić. Interactive high-quality green-screen keying via color unmixing. *ACM Trans. Graph.*, 36(4), August 2016.

[2] Yağiz Aksoy, Tunç Ozan Aydin, Aljoša Smolić, and Marc Pollefeys. Unmixing-based soft color segmentation for image manipulation. *ACM Trans. Graph.*, 36(4), March 2017.

[3] Yağız Aksoy, Tae-Hyun Oh, Sylvain Paris, Marc Pollefeys, and Wojciech Matusik. Semantic soft segmentation. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 37(4):72:1–72:13, 2018.

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:2481–2495, 2017.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *CoRR*, abs/1412.7062, 2014.

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:834–848, 2016.

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *ArXiv*, abs/1706.05587, 2017.

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[10] Jose Dolz, Karthik Gopinath, Jing Yuan, Herve Lombaert, Christian Desrosiers, and Ismail Ben Ayed. Hyperdense-net: A hyper-densely connected cnn for multi-modal image segmentation. *IEEE Transactions on Medical Imaging*, 38: 1116–1126, 2018.

[11] Nanqing Dong and Eric P. Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, 2018.

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[13] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6602–6611, 2016.

[14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *ArXiv*, abs/1406.2661, 2014.

[15] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, 2016.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[17] Qi-Xing Huang and Leonidas Guibas. Consistent shape maps via semidefinite programming. In *Proceedings of the Eleventh Eurographics/ACMSIGGRAPH Symposium on Geometry Processing*, SGP '13, 2013.

[18] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[19] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew C. H. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *ArXiv*, abs/1804.03999, 2018.

[20] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alexei A. Efros, and Sergey Levine. Conditional networks for few-shot semantic segmentation. In *ICLR*, 2018.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[23] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *ArXiv*, abs/1610.02391, 2016.

[24] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In Gabriel Brostow Tae-Kyun Kim, Stefanos Zafeiriou and Krystian Mikolajczyk, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 167.1–167.13. BMVA Press, September 2017.

[25] Mennatullah Siam and Boris N. Oreshkin. Adaptive masked weight imprinting for few-shot segmentation. *CoRR*, abs/1902.11123, 2019.

[26] Dheeraj Singaraju and René Vidal. Estimation of alpha mattes for multiple image layers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33:1295 – 1309, 08 2011.

[27] Ashish Sinha and José Estevan Dolz. Multi-scale guided attention for medical image segmentation. *ArXiv*, abs/1906.02849, 2019.

[28] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. 03 2017.

[29] Yu-Wing Tai, Jiaya Jia, and Chi-Keung Tang. Soft color segmentation and its applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(9), September 2007.

[30] Jianchao Tan, Jyh-Ming Lien, and Yotam Gingold. Decomposing images into layers via rgb-space geometry. *ACM Trans. Graph.*, 36(1), November 2016.

[31] Fan Wang, Qixing Huang, and Leonidas J. Guibas. Image cosegmentation via consistent functional maps. In *Proceedings of the 2013 IEEE International Conference on Computer Vision*, ICCV '13, 2013.

[32] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, D. Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *ArXiv*, abs/1908.07919, 2019.

[33] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. *ArXiv*, abs/1908.06391, 2019.

[34] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. pages 2868–2876, 10 2017. doi: 10.1109/ICCV.2017.310.

[35] Christopher Zach, Manfred Klopschitz, and Manfred Pollefeys. Disambiguating visual relations using loop constraints. pages 1426–1433, 06 2010. doi: 10.1109/CVPR.2010.5539801.

[36] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[37] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. 10 2018.

[38] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, 2016.

[39] Tinghui Zhou, Yong Jae Lee, Stella X. Yu, and Alexei A. Efros. Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In *CVPR*, pages 1191–1200. IEEE Computer Society, 2015.

[40] Tinghui Zhou, Philipp Krähenbühl, Mathieu Aubry, Qixing Huang, and Alexei Efros. Learning dense correspondence via 3d-guided cycle consistency. 04 2016.

[41] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.