GuidedMix-Net: Semi-Supervised Semantic Segmentation by Using Labeled Images as Reference

Peng Tu[*1, 2], **Yawen Huang**[*3], **Zheng Feng**[†1], **Zhenyu He**[4], **Liujuan Cao**[5], **Ling Shao**[6]

[1]Southern University of Science and Technolog, Shenzhen, China   [2]Shenzhen Microbt Electronics Technology Co., Ltd, China
[3]Tencent Jarvis Lab, Shenzhen, China   [4]Harbin Institute of Technology, Shenzhen, China   [5]Xiamen University, Xiamen, China
[6]National Center for Artificial Intelligence, Saudi Data and Artificial Intelligence Authority, Riyadh, Saudi Arabia
yh.peng.tu@gmail.com, yawenhuang@tencent.com, zfeng02@gmail.com
zhenyuhe@hit.edu.cn, caoliujuan@xmu.edu.cn, ling.shao@ieee.org

Abstract

Semi-supervised learning is a challenging problem which aims to construct a model by learning from limited labeled examples. Numerous methods for this task focus on utilizing the predictions of unlabeled instances consistency alone to regularize networks. However, treating labeled and unlabeled data separately often leads to the discarding of mass prior knowledge learned from the labeled examples. In this paper, we propose a novel method for semi-supervised semantic segmentation named GuidedMix-Net, by leveraging labeled information to guide the learning of unlabeled instances. Specifically, GuidedMix-Net employs three operations: 1) interpolation of similar labeled-unlabeled image pairs; 2) transfer of mutual information; 3) generalization of pseudo masks. It enables segmentation models can learning the higher-quality pseudo masks of unlabeled data by transfer the knowledge from labeled samples to unlabeled data. Along with supervised learning for labeled data, the prediction of unlabeled data is jointly learned with the generated pseudo masks from the mixed data. Extensive experiments on PASCAL VOC 2012, and Cityscapes demonstrate the effectiveness of our GuidedMix-Net, which achieves competitive segmentation accuracy and significantly improves the mIoU by +7% compared to previous approaches.

1   Introduction

The past several years have witnessed the success of convolutional neural networks (CNNs) (Long, Shelhamer, and Darrell 2015; Ronneberger, Fischer, and Brox 2015b; Huang et al. 2020; Chen et al. 2017; Huang et al. 2021) for visual semantic segmentation. Although data-driven deep learning techniques have benefitted greatly from the availability of large-scale image datasets, they require dense and precise pixel-level annotations for parameter learning. Alternative learning strategies, such as semi-supervision, have thus emerged as promising approaches to reduce the need for annotations, requiring simpler or fewer labels for image classification (Tarvainen and Valpola 2017a; Takeru et al. 2018; Sohn et al. 2020).

[*]Equal Contribution
[†]Corresponding Author

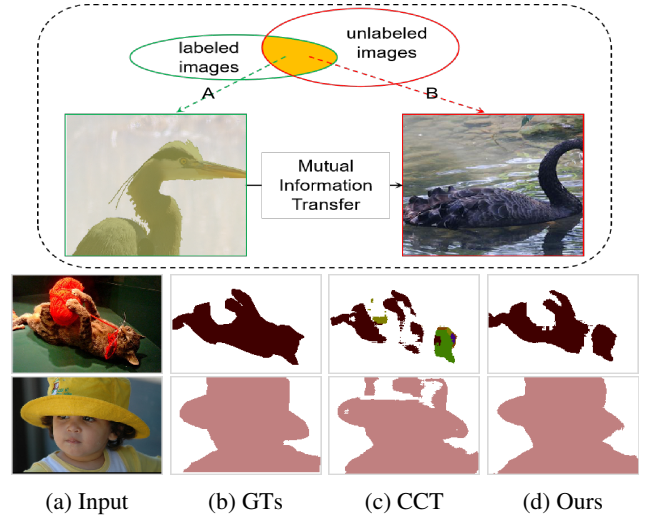(a) Input   (b) GTs   (c) CCT   (d) Ours

Figure 1: **Upper** (dotted box): The mutual information transfer module selects similar features to transfer knowledge from the labeled samples to unlabeled images. **Lower**: Examples of ground-truths (GTs) (b), the pseudo mask of CCT (c) and ours (d).

Recent semi-supervised methods for semantic segmentation, such as (Ouali, Hudelot, and Tami 2020; Luo et al. 2020; French et al. 2020; Olsson et al. 2021), exploit consistency by perturbing the unlabeled samples to regularize model training. Intuitively, the models are expected to manifest the invariance underlying any small perturbations while observing the natural properties of the data, especially for unlabeled data. To address this, numerous semi-supervised semantic segmentation methods have been proposed against the perturbations when leveraging unlabeled samples. For example, CCT (Ouali, Hudelot, and Tami 2020) introduces random manual perturbations by designing separated, unrelated decoders for each type of perturbation. DTC (Luo et al. 2020) builds a task-level regularization rather than data-level perturbation. CutMix (French et al. 2020) and Class-Mix (Olsson et al. 2021) follow MixUp (Zhang et al. 2017) and achieve semi-supervised segmentation by forcing the predictions for the augmented and original data to be consistent.

Although various approaches have been introduced over the years, a key bottleneck of semi-supervised segmentation models is that they typically treat the labeled and unlabeled samples separately during training. Existing methods focus on how to use unlabeled data alone under various manual perturbations. On the one hand, although low-level perturbations do improve the robustness slightly, the rich underlying intrinsic information of the unlabeled instances has not yet been fully explored. For instance, consistency-based methods rely primarily on the local information of the samples themselves by constraining the local smoothness. This strategy cannot comprehensively mine the structural information, especially for unlabeled data, which in turn causes the model to produce a suboptimal solution. Another clear weakness of consistency-based methods is that they require abundant and diverse perturbations, which are expensive and time-consuming to obtain. For instance, CCT (Ouali, Hudelot, and Tami 2020) incorporates almost thirty decoders to make the learned model adequately robust. The example results shown in Fig. 1 (c) also confirm that generating pseudo masks using simple perturbation consistency training yields significant deviation in the contour and semantic understanding of objects. In other words, the massive amount of prior information learned from the labeled samples cannot be transferred to the unlabeled data. Current semi-supervised semantic segmentation methods provide inconsistent optimization objectives for labeled and unlabeled data in different training stages, where the labeled samples are used to improve the discriminative ability and the unlabeled samples enhance the smoothness of models. However, we should be able to train a uniform model by leveraging a large amount of unlabeled data under the guidance of labeled samples. This will enable the learned representations to be refine, thus facilitating mutual information interaction and transfer. We also note that humans can recognize unfamiliar objects subconsciously, by making inferences based on similar or recognizable objects. For example, in Fig. 1, the dotted box provides a labeled image A and an unlabeled image B with similar objects. Most people can recognize and segment the object in image B by transferring their knowledge of image A to image B. In contrast, existing deep models are typically trained with limited labeled samples and most directly generate the pseudo mask of image B, making it difficult to produce a high-quality prediction for unseen samples, as shown in the bottom of Fig. 1 (c). The example shown in Fig. 1 is a relatively simple scenario; natural images are usually far more complex with, for example, multiple occluded objects, making them even more challenging to segment. Another observation is that similar objects (*e.g.* intra-class objects) often contain common edges and textures. An intuitive way to improve the segmentation of unlabeled data is therefoce to refer to labeled images, as humans do.

Motivated by these problems, we propose a novel semi-supervised method for semantic segmentation, named GuidedMix-Net. GuidedMix-Net allows knowledge to be transferred from the labeled images to the unlabeled samples, as occur in the human cognitive path. To learn from the unlabeled samples, GuidedMix-Net employs three processes, *i.e.*, labeled-unlabeled image pair interpolation, mu-

tual information transfer, and pseudo mask generation. Specifically, we feed pairs of labeled and unlabeled images as input into the model and carry out a linear interpolation of them to capture pairwise interactions. Then, we learn the uniform feature vectors from the mixed data to inherit different contexts from the image pairs. To incorporate non-local blocks (Wang et al. 2018) into the mixed feature layer, long-range dependencies are explored both within images and between them to mine similar object patterns and learn semantic correlations. We further select objects with similar features to ensure that the cues will be similar for different image pairs. Feature selection improves the prediction and mask qualities of the unlabeled images by using the supervised information from the labeled images as reference. After that, we decouple the hybrid prediction to obtain a pseudo mask for the unlabeled image. As a result, the generated pseudo masks are more credible than the direct predictions of unlabeled samples. Finally, the pairs can be utilized for self-training to explore the rich underlying semantic structures provided by the unlabeled examples and further improve the performance of our model.

2 Related Work

2.1 Semi-Supervised Classification

Semi-supervised classification methods (Sajjadi, Javanmardi, and Tasdizen 2016; Tarvainen and Valpola 2017b; Takeru et al. 2018) typically focus on achieving consistent training by combining a standard supervised loss (*e.g.* cross-entropy loss) and an unsupervised consistency loss to encourage consistent predictions for perturbations on the unlabeled samples. Randomness is essential for machine learning to either guarantee the generalization and robustness of the model or provides multiple different predictions for the same input. Based on this, Sajjadi et al. (Sajjadi, Javanmardi, and Tasdizen 2016) introduced an unsupervised loss function which leverages the stochastic property of randomized data augmentation, dropout and random max-pooling to minimize the difference between the predictions of multiple passes of a training sample through the network. Although these random augmentation techniques can improve the performance, they still remain difficulty on providing effective constraints for boundaries. Miyato et al. (Takeru et al. 2018) after proposed a virtual adversarial training scheme to achieve smooth regularization. Their method aims to perturb the decision boundary of the model via a virtual adversarial loss-based regularization to measure the local smoothness of the conditional label distribution.

2.2 Semi-Supervised Semantic Segmentation

Semi-supervised semantic segmentation algorithms have achieved great success in recent years (Papandreou et al. 2016; Wei et al. 2018; Souly, Spampinato, and Shah 2017; Lee et al. 2019; Hung et al. 2018). For example, EM-Fixed (Papandreou et al. 2016) provides a novel online expectation-maximization method by training from either weakly annotated data such as bounding boxes, image-level labels, or a combination of a few strongly labeled and many

weakly labeled images, sourced from different datasets. EM-Fixed benefits from the use of both a small amount of labeled and large amount of unlabeled data, achieving competitive results even against other fully supervised methods. From a certain point of view, EM-Fixed involves weak labels for training and hence is not a pure semi-supervised method. Spampinato et al. (Souly, Spampinato, and Shah 2017) designed a semi-supervised semantic segmentation method using limited labeled data and abundant unlabeled data in a generative adversarial network (GAN). Their model uses discriminators to estimate the quality of predictions for unlabeled data. If the quality score is high, the pseudo-label generated from the prediction can be regarded as the ground-truth, and the model is optimized by calculating the cross-entropy loss. However, models trained on a limited amount of labeled data typically fail in the following ways: 1) they generate inaccurate low-level details; 2) they misinterpret high-level information. To address these problems, s4GAN-MLMT (Mittal, Tatarchenko, and Brox 2019) fuses a GAN-based branch and a classifier to discriminate the generated segmentation maps. However, considering the intrinsic training difficulty of GANs, some semi-supervised image classification approaches instead adopt a consistent training strategy to ensure similar outputs under small changes, since this is flexible and easy to implement. CCT (Ouali, Hudelot, and Tami 2020) employs such a training scheme for semi-supervised semantic segmentation, where the invariance of the predictions is enforced over different perturbations applied to the outputs of the decoder. Specifically, a shared encoder and a main decoder are trained in a supervised manner using a few labeled examples. To leverage unlabeled data, CCT enforces consistency between the main decoder predictions and the other set of decoders (for each type of perturbations), and uses different perturbations output from the encoders as input to improve the representations.

Unlike previous methods, which primarily focus on learning unlabeled data, we use the labeled images as references and transfer their knowledge to guide the learning of effective information from unlabeled data. As a result, the proposed method generate high-quality features from the labeled images, and refines features of unlabeled data via pairwise interaction.

3 GuidedMix-Net

Assume that we have a limited number of labeled images $S_l=\{x_l, y_l\}$, where y_l is the ground-truth mask of the image x_l, and a large amount of data without annotations $S_u=\{x_u\}$. The image $x \in \mathbb{R}^{H \times W}$ has spatial dimensions of $H \times W$ and the masks $y \in \mathbb{R}^{H \times W \times C}$ have C categories. Fully supervised methods aim to train a CNN $\Gamma(x; \theta)$ that takes image x as input, where θ denotes the parameter of the model, and outputs the segmented mask $\hat{y}$ by minimizing the cross-entropy loss $\mathbb{L}_{ce}$ as follows:

$$\mathbb{L}_{ce}(\hat{y}, y) = -\sum_i \hat{y}_i log(y_i), \qquad (1)$$

where i represents the i-th category. Generally, collecting large-scale labeled training data is time-consuming, costly

and sometimes infeasible. In contrast, for some computer vision tasks, large amounts of unlabeled data can be collected relatively easily and labeled samples are hard to be obtained. In this case, fully supervised training scheme cannot achieve good performance when suffering in the presence of a minor data deficiency. To address this and employ unlabeled examples during training, we propose a novel framework, called GuidedMix-Net, to leverage the limited number of labeled samples to guide the learning of unlabeled data. The overall framework is shown in Fig. 2.

To leverage the labeled samples to guide the generation of credible pseudo masks for the unlabeled samples, GuidedMix-Net employs three operations: 1) interpolation of image pairs; 2) transfer of mutual information; 3) generation of pseudo masks, which will be introduced accordingly.

3.1 Labeled-Unlabeled Image Pair Interpolation

Labeled-unlabeled image pair interpolation (LUPI) applies linear interpolation to formulate a data mixing objective for current unlabeled instances with potentially similar labeled samples to guarantee the cues to be similar, and enable information to flow between them. Given a pair of samples (x_l^i, y_l^i) and x_u^k, image-level's interpolation as shown in Eq. 2.

$$x_{mix}(x_l, x_u) = \lambda x_l + (1 - \lambda)x_u. \qquad (2)$$

The output after images interpolation can be expressed as x_{mix}. To learn unlabeled data x_u over the labeled samples x_l, we set $\lambda \leftarrow \min(\lambda, 1 - \lambda)$, where $\lambda \in (0, 1)$ is a hyper-parameter sampled from the $Beta(\alpha, \alpha)$ distribution with α. And here the α is predefined as 1.

Similar Image Pair Selection Although LUPI enables models to associate similar cues between the labeled and unlabeled images, randomly selecting an image pair will not allow knowledge to be transferred from labeled to unlabeled data, since the individual pair may not contain enough similar objects. We overcome this problem by construct similar image pairs in a mini-batch for training. Specifically, we add a fully connected layer as a classifier after the encoder to enhance the semantics of the pooled features, and select image pairs with similar features according to the Euclidean distance (shown in Eq. 3, where Γ_{enc} is the encoder module of GuideMix-Net), to conduct image pairs. Note that the classifier is first trained using labeled data, so it has some recognition ability. This procedure allows the proposed model to capture the most similar labeled example for each unlabeled image. Further discussion can be see in Appendices A.

$$d(x_l^i, x_u^k) = \sqrt{\sum^H \sum^W (\Gamma_{enc}(x_l^i) - \Gamma_{enc}(x_u^k))^2}. \qquad (3)$$

3.2 Mutual Information Transfer

After mixing the pair of samples, we associate similar cues to enhance the features and generate pseudo masks of the unlabeled samples. Generally, labeled data corresponds to the credible features, and unlabeled data are treated as poor-quality features, since no supervision signals are present to guide the gradient updates. This means that in a uniform
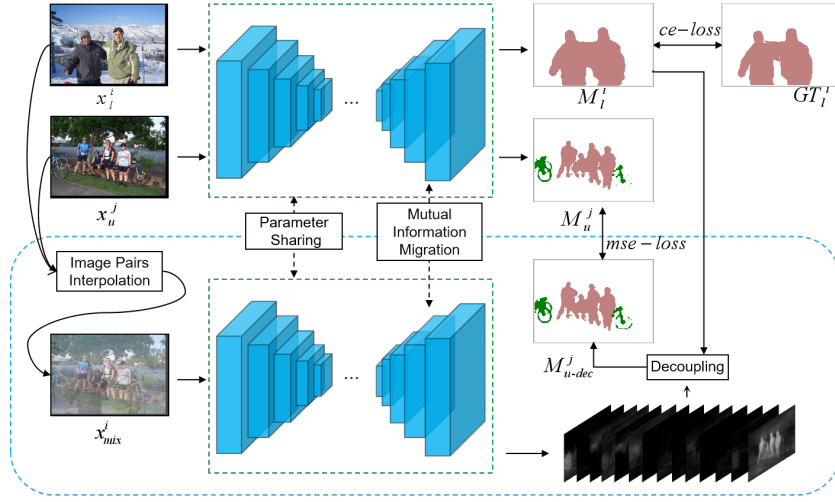
Figure 2: Overview of our proposed semi-supervised segmentation approach. GuidedMix-Net follows the basic architecture of U-Net, consisting of an encoder-decoder architecture. The main decoder is constructed by ResNet, while the decoder is incorporated by our MITrans modules.
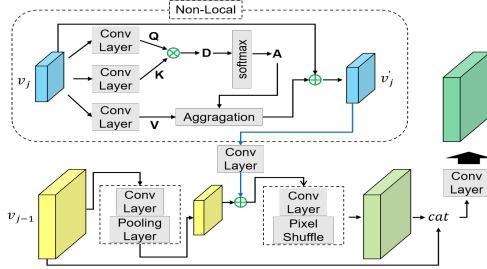

Figure 3: Construction details of the MITrans module.

mixed vector space, for a pair of similar objects from different sources, the poor features can use the credible ones as reference to improve their quality, whether the credible information is located in the short- or long-range. Although LUPI enables the short-range credible information to flow to the worse features (see the Appendices A), CNNs are ineffective in capturing information from distant spatial locations. We solve this problem by applying a non-local (NL) block (Ouali, Hudelot, and Tami 2020) to obtain long-range patches that are similar to a given local region of mixed data. After mixing labeled-unlabeled images, we obtain a mutual image which contains all the information from the input pairs. The mixed data x_{mix} is then fed to the encoder for segmentation Γ until reaching layer j, providing an intermediate $\mathbf{v}_j$ as follows:

$$\mathbf{v}_j = h_j(x_{mix}). \qquad (4)$$

Another intermediate, $\mathbf{v}_{j-1}$, is produced by the $j-1$ layer, where the spatial size of which is double that of $\mathbf{v}_j$. The non-local module collects contextual information from long-range features to enhance local feature representation. As shown in the dotted box of Fig. 3, the module uses two convolutional layers of 1×1 and filters $\mathbf{v}_j$ to map and obtain two features $\mathbf{Q}$ and $\mathbf{K}$. The spatial size is the same as

for $\mathbf{v}_j$, but the channel number is half that of $\mathbf{v}_j$ double for dimension reduction. Then, we generate a correlation matrix $\mathbf{D}$ by calculating the correlation between $\mathbf{Q}$ and $\mathbf{K}$. A softmax layer is also utilized on $\mathbf{D}$ over the channel dimension to get the attention map $\mathbf{A} = f(\mathbf{Q}, \mathbf{K})$. To obtain the adapted features of $\mathbf{v}_j$, another convolutional layer with a 1×1 filter is used to generate $\mathbf{V}$ without size change. The long-range contextual information is captured by an aggregation operation, as follows:

$$\mathbf{v}'_{j,n} = \frac{1}{C(x)} \sum_{\forall \mathbf{m}} f(\mathbf{Q}_n, \mathbf{K}_m)\mathbf{V_m} + \mathbf{v}_{j,n}, \qquad (5)$$

where n and m are the indices of the position in the variable space whose response needs to be computed and all other potential positions, respectively. Parameters $\mathbf{v}'_{j,n}$ denotes the feature vector in the final output features $\mathbf{v}'_j$ at position $\mathbf{n}$, and $\mathbf{v}_{j,n}$ is a feature vector in $\mathbf{v}_j$ at position $\mathbf{n}$. The function f is used to represent relationships such as the affinity between $\mathbf{Q}_n$ and all $\mathbf{K}_m$. Finally, $C(x)$ is a normalization factor.

After that, several convolutional layers are combined with the PixelShuffle layer (Shi et al. 2016) to fuse the low-level features and restore the spatial information (shown in Fig. 3). The model is first trained on the labeled samples before mixing. Although the number of training samples is small, they provide some recognition ability. The non-local blocks use the features of the labeled samples as the novel training signals to correct the feature generation of the unlabeled samples in the mixed data (i.e. inter-images). Unlike previous methods, which primarily focus on intra-image information, the "mutual information transfer" module addresses the semantic relations within for images comprehensive object pattern mining. The proposed module captures semantic similarities for image themselves to build a mutual information transformation mode, and thus improves the prediction of the unlabeled samples.

Note that the non-local blocks in MITrans are used to associate similar patches from the labeled and unlabeled features. Therefore, they can leverage supervision signals for unlabeled data training.

3.3 Pseudo Mask Generation

To effectively learn from the unlabeled samples, we need to decouple high-quality pseudo masks from the mixed data x_{mix}. According to the translation equivariance (Goodfellow, Bengio, and Courville 2015) of the convolution operator, the translation operated on the input image is still detectable on the output features with the corresponding translation. The translation equivariance can be reflected in the mixed data as well, as shown in Appendices B, the visualization (a), (b), (c), where the activations for the spatial locations of the interesting object, $e.g.$, bus and person, are invariant. Further, the predicted layer of the segmentation network assigns these activated features to certain category channels, separately. We focus on this and propose a pseudo mask generation (PMG) to generate the masks for the unlabeled instances by conducting subtraction between the predictions to constrain foreground separation, which comes from labeled and unlabeled data, respectively. We then train both the labeled and unlabeled data jointly.

In general, semantic segmentation can be regarded as seeking a mapping function Γ, such that the output $M = \Gamma(x)$ is the desired mask which is close to the ground-truth. For a pair of labeled and unlabeled images (x_l, x_u), the predictions are $M_l = \Gamma(x_l)$ and $M_u = \Gamma(x_u)$, respectively. After feeding x_{mix} from Eq. 2 into the segmentation network Γ, we can obtain the predicted mask $M_{mix} = \Gamma(x_{mix})$, which can be treated as the approximation of directly mixing the masks M_l and M_u:

$$M_{mix} = \Gamma(x_{mix}) \approx M_l + M_u. \quad (6)$$

We decouple the pseudo masks of x_u from M_{mix}, and then leverage them as the target to calculate the mean squared error loss (with the direct output of x_u from the main decoder). This procedure ensures the model to be robust and less sensitive to small perturbations.

Hard Decoupling The goal of mask decoupling is to eliminate M_l from the mixed data and then generate pseudo masks for the unlabeled samples. Considering that the ground-truth of labeled data are provided for the model in the early training stage, the prediction M_l has higher probability of being close to the real mask. Once M_l is obtained, we can directly decouple the unlabeled data mask M_{u-dec} using Eq. 7, which we refer to as hard decoupling:

$$M_{u-dec} = M_{mix} - M_l. \quad (7)$$

Hard decoupling is reasonable since the neural network has the ability to separate the corresponding category channels (an example is shown in Appendices B). Directly subtracting between final predictions can separate and obtain more refined results for the unlabeled samples.

Soft Decoupling The proposed hard decoupling directly performs a subtraction between the prediction of the mixed image and the labeled image, which may counteract the prediction of the overlapping region of objects in the mixed image. To overcome this problem, we propose a soft decoupling for the pseudo mask generation as follows:

$$M_{u-dec} = M_{mix} - \lambda M_l, \quad (8)$$

where λ is the parameter from the $Beta(\alpha, \alpha)$ distribution. Soft decoupling retains the details of the overlapping region by weakening the intensity of M_l in M_{mix}. As shown in Table 2, soft decoupling is better than hard decoupling.

GuidedMix-Net pays attention to the outline of objects in a complex environment by first transferring knowledge from the mixed labeled-unlabeled pairs, and then decoupling their predictions, to understand the complete semantic information of objects. As shown in Appendices D, GuidedMix-Net is complements object contours and semantic understanding.

3.4 Loss Function

We develop an overall loss function $\mathbb{L}$ for our consistency based semi-supervised learning (SSL) as follows:

$$\mathbb{L} = \mathbb{L}_{sup} + \omega_{usup}\mathbb{L}_{usup}, \quad (9)$$

where ω_{usup} is an unsupervised loss weight, such as (Laine and Aila 2016), that controls the balance between the two losses. On the one hand, $\mathbb{L}_{usup}$ in Eq. 10 is an unsupervised mean squared error (MSE) loss to calculate the difference between the decoupling mask M_{u-dec} and the direct prediction M_u:

$$\mathbb{L}_{usup} = \frac{1}{H*W} \sum^{H} \sum^{W} (M_{u-dec} - M_u)^2. \quad (10)$$

On the other hand, for supervised training, the loss $\mathbb{L}_{sup}$ consists of three terms to optimize the model as follows:

$$\mathbb{L}_{sup} = L_{ce}(M_l, y_l) + L_{dec} + L_{cla}, \quad (11)$$

where $L_{ce}(M_l, y_l)$ is the same as in Eq. 1, and L_{cla} is the classifier loss term for image-level annotations. For L_{dec}, we first select a sample $\hat{x}_l$ for a seed x_l according to the matching rules in Sec. 3.1, where λ follows Sec. 3.1. We then denote $\hat{y}_l$ and $\hat{M}_l$ (y_l and M_l) as the corresponding ground-truth and prediction of $\hat{x}_l$ (the seed x_l), respectively. In addition, a mixup operation can be conducted on both labeled sample x_l and $\hat{x}_l$ following Eq. 2 to obtain a mixed sample $x^l_{mix} = \lambda x_l + (1 - \lambda)\hat{x}_l$. A consistency loss between decoupled masks $M_{dec} = M_{mix} - M_l$ and the prediction $\hat{M}_l$ can be defined as

$$\mathbb{L}_{dec} = \frac{1}{H*W} \sum^{H} \sum^{W} (M_{dec} - M_l)^2. \quad (12)$$

Another analysis of GuidedMix-Net shown in Appendices C.

4 Experiments

4.1 Dataset and Evaluation Metrics

PASCAL VOC 2012. This dataset is widely used for semantic segmentation and object detection. It consists of 21

classes including background. We use 1,464 training images and 1,449 validation images from the original PASCAL dataset, and also leverage the augmented annotation dataset (involving 9,118 images) (Hariharan et al. 2011) like (Huang et al. 2018; Zhao et al. 2017).

Cityscapes. We use Cityscapes to further evaluate our model. This dataset provides different driving scenes distributed in 19 classes, with 2,975, 500, 1,525 densely annotated images for training, validation and testing. For semantic segmentation, 59 semantic classes and 1 background class are used in training and validation, respectively.

Evaluation Metric. Common data augmentation methods are used during our training procedure, which including random resizing (scale: 0.5~2.0), cropping (321×321 for PASCAL VOC 2012, 513×513 for Cityscapes, 480×480 for PASCAL Context), horizontal flipping and slight rotation. We evaluate different methods by measuring the averaged pixel intersection-over-union (IoU).

4.2 Network Architecture and Training Details

Encoder. The encoder is based on ResNet (He et al. 2016) pretrained on ImageNet (Krizhevsky, Sutskever, and Hinton 2012), and also includes the PSP module (Zhao et al. 2017) after the last layer.

Decoder. GuidedMix-Net combines labeled and unlabeled data by image linear interpolation, setting the new SOTA, for semi-supervised semantic segmentation. To avoid the disintegrates of details for the mixed pairs, we employ a skip connection in the decoder, as done in U-Net (Ronneberger, Fischer, and Brox 2015a). A pixel shuffle layer (Shi et al. 2016) is also utilized to restore the spatial resolution of features.

Training Details. Similar to (Chen et al. 2017), we use a "poly" learning rate policy, where the base learning rate is multiplied by $((1 - \frac{iter}{max_iter})power)$ and $power = 0.9$. Our segmentation network is optimized using the stochastic gradient descent (SGD) optimizer with a base learning rate of 1e-3, momentum of 0.9 and a weight decay of 1e-4. The model is trained over 40,000 iterations for all datasets, and the batch-size is set to 12 for PASCAL VOC 2012, and 8 for Cityscapes and PASCAL Context. We conduct all our experiments on a Tesla V-100s GPU.

4.3 Results on Pascal VOC 2012

Ablation Studies Our ablation studies examine the effect of different values of λ and the impact of different components in our framework.

Different λ Values. The results under different values of λ are reported in Table 1. An can be see, changing λ used in Sec. 3.1 impacts the results, because λ controls the intensity of pixels in the mixed input data. As shown in Table 1, too high or too low a λ is not conducive to model optimization. A high λ value leads to the labeled information being discarded, while a low λ value results in the unlabeled data being covered. When $\lambda < 0.5$, GuidedMix-Net provides the best performance on PASCAL VOC 2012. We thus select $\lambda < 0.5$ for the remaining experiments on PASCAL VOC 2012.

Table 1: The impact of different lambda values on the experimental results of PASCAL VOC 2012.

λ	backbone	Data		mIoU
		labels	unlabels	
< 0.1				67.9
< 0.2				69.5
< 0.3	ResNet50	1464	9118	70.7
< 0.4				71.4
< 0.5				**73.7**

Table 2: Ablation studies of using similar image pairs, MITrans, and hard & soft decoupling modules in GuidedMix-Net. We train the models on ResNet50 and test them on the validation set of PASCAL VOC 2012.

Method	used	Data		mIoU
		labels	unlabels	
Similar Pair	×			71.9
	√			**73.7**
MITrans	×	1464	9118	72.7
	√			**73.7**
Hard Decoupling	√			71.4
Soft Decoupling	√			**73.7**
Suponly w/o MITrans	√	1464	9118	70.2
Suponly $w/$ MITrans	√			70.5

Different Components. As shown in Table 2, we evaluate the influence of different components of GuidedMix-Net. For fair comparison, we evaluate one component per experiment and freeze the others. Firstly, we investigate different strategies for constructing image pairs, *i.e.*, random selection of similar pairs. i) The first strategy randomly selects a labeled image for each unlabeled image in a mini-batch. ii) To seek similar pairs of labeled and unlabeled images, we add a classifier after the encoder model. We match the most similar unlabeled images for each labeled sample according to the Euclidean distance between the features. As shown in the Table 2, the similar pairs bring a 2.5% mIoU gain over the plain random selection (71.9% vs. 73.7%). The construction of similar pairs provides context for the target objects in the subsequent segmentation task, and assists GuidedMix-Net in transferring knowledge from the labeled images to the unlabeled samples, with little increasing complexity. Secondly, we explore whether MITrans is useful for knowledge transfer. The results provided in the Table 2 clearly show that MITrans achieves a significant mIoU gain of 1.4% (72.7% vs. 73.7%) by explicitly referencing similar and high-confidence non-local feature patches to refine the coarse features of the unlabeled samples. Thirdly, as shown in the Table 2, the performance of soft decoupling is 3.2% better than hard decoupling (71.4% vs. 73.7%), since soft decoupling considers the overlapping that occurs in the mixed data and tends to preserve local details. The various components used in our GuidedMix-Net are beneficial alone, and therefore combining them leads to significantly improved optimization.

The above experimental results indicated that all of our designed components were conducive to semi-supervisedd

semantic segmentation learning. However, we still puzzle with was the performance gain of the MITrans module attributed to the knowledge transfer of using labeled samples as a reference paradigm when learning unlabeled data representations? Or to the feature expression of labeled samples enhanced by nonlocal blocks? We further examined this question by additional experiments, as shown in Table 2. "Suponly w/o MITrans" was the experiment training on the labeled samples only and the nonlocal blocks in MITrans were removed to obtain 70.2 mIoU. IN contrast, "Suponly $w/$ MITrans" was the experiment training on the labeled samples only, and MITrans was fully used to obtain 70.5 mIoU. This set of experiments clearly showed that there was little performance gain for nonlocal labeled samples. The huge performance gains described above are mainly due to MITrans' ability to effectively transfer knowledge from labeled samples to unlabeled data.

Semi-Supervised Semantic Segmentation The experimental setting of CCT (Ouali, Hudelot, and Tami 2020) is different from other semi-supervised semantic segmentation methods AdvSSL (Hung et al. 2018), S4L (Zhai et al. 2019), GCT (Ke et al. 2020), CutMix (French et al. 2020), Reco (Liu et al. 2021), ClassMix (Olsson et al. 2021) and SSContrast (Alonso et al. 2021). Here, we provide the comparison experiment in Table 3. We can see that GuidedMix-Net outperforms CCT with a performance increase of over 5.7%.

Comparing with Other State of the Arts. We explore the performance using the deeper backbone of ResNet101 for the semi-supervised semantic segmentation task. The results show in Table 3. GuidedMix-Net outperforms the current methods for semi-supervised image segmentation by 3.4%, 3.7%, and 3.4% on 1/8 labels, 1/4 labels, 1/2 labels, respectively. The significant performance gains on different ratios of labeled data demonstrate that GuidedMix-Net is a generally efficient and effective semi-supervised semantic segmentation method. Quality visualization results provided in Appendices D.

Table 3: Comparison with other state-of-the-art semi-supervised semantic segmentation methods under different ratios of labeled data on PASCAL VOC 2012.

SSL	1/8	1/4	1/2	-
AdvSSL	68.4	70.8	73.3	-
S4L	67.2	68.4	72.0	-
GCT	70.7	72.8	74.0	-
ReCo	71.0	-	-	-
CutMix	70.8	71.7	73.9	-
Ours	**73.4**	**75.5**	**76.5**	-

SSL	500 labels	1000 labels	1464 labels	Memory Size
CCT	58.6	64.4	69.4	24kM
Ours	**65.4**	**68.1**	**73.7**	**15kM**

4.4 Results on Cityscapes

Cityscapes has 2,975 training images. In our experiments, we divide them into 1/8 labels and 1/4 labels, while the remaining data are treated as unlabeled. We use ResNet101 as

Table 4: Influence of different lambda values on the experimental results of Cityscapes.

λ	backbone	Data		mIoU
		labels	unlabels	
< 0.1				64.3
< 0.2				64.0
< 0.3	ResNet101	1/8	7/8	**65.8**
< 0.4				65.5
< 0.5				65.7

Table 5: Comparison with other semi-supervised semantic segmentation methods under different ratios of labeled data on Cityscapes.

SSL Methods	100 labels	1/8 labels	1/4 labels	1/2 labels
AdvSSL	-	57.1	60.5	-
s4GAN	-	59.3	61.9	-
CutMix	51.2	60.3	63.9	-
ClassMix	54.1	61.4	63.6	66.3
ReCo	56.5	64.9	67.5	68.7
Ours	**56.9**	**65.8**	**67.5**	**69.8**

the backbone to train the models. Since the optimal value of λ varies with the training dataset, we conduct experiments on Cityscapes leveraging 1/8 labeled images as the training data to explore the impact of λ on this dataset, and show the results in Table 4. For Cityscapes, when $\lambda <$ 0.3, GuidedMix-Net achieves 65.8 mIoU on the validation dataset, which is better than other selected value ranges. We thus fix the value of λ to be less than 0.3, and verify the gap between GuidedMix-Net and other approaches. Relevant results are presented in Table 5. GuidedMix-Net yields considerable improvements on Cityscapes over other semi-supervised semantic segmentation methods, *i.e.*, mIoU increases of 0.7%, 1.4%, and 1.6% for the 100, 1/8, and 1/2 labels, respectively. The distribution of different classes on Cityscapes is highly imbalanced. The vast majority of classes are present in almost every image, and the few remaining classes occur scarcely. As such, inserting a classifier after the encoder to semantically enhance features and assist in matching similar images is unhelpful. Thus, we use the mixture of randomly selected image pairs in GuidedMix-Net. We also provide the visul results on Cityscapes in Appendices E using only 1/8 labeled images.

In addition, the experiments of GuidedMix-Net, which compares with other SOTA on PASCAL-Context dataset had introduced in Appendices F.

5 Conclusion

This paper has presented a novel semi-supervised learning method for semantic segmentation, called GuidedMix-Net, and achieves SOTA performance. In the future, we will investigate the use of unlabeled data in other related areas, such as medical imaging. We will continue improving the learning mechanism of the unlabeled samples guided by labeled data.

References

[Alonso et al. 2021] Alonso, I.; Sabater, A.; Ferstl, D.; Montesano, L.; and Murillo, A. C. 2021. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. *arXiv preprint arXiv:2104.13415*.

[Chen et al. 2017] Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40(4):834–848.

[French et al. 2020] French, G.; Laine, S.; Aila, T.; Mackiewicz, M.; and Finlayson, G. 2020. Semi-supervised semantic segmentation needs strong, varied perturbations. In *British Machine Vision Conference*, number 31.

[Goodfellow, Bengio, and Courville 2015] Goodfellow, I.; Bengio, Y.; and Courville, A. C. 2015. Deep learning. *Nature* 521:436–444.

[Hariharan et al. 2011] Hariharan, B.; Arbelaez, P.; Bourdev, L. D.; Maji, S.; and Malik, J. 2011. Semantic contours from inverse detectors. In *International Conference on Computer Vision*.

[He et al. 2016] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

[Huang et al. 2018] Huang, Z.; Wang, X.; Wang, J.; Liu, W.; and Wang, J. 2018. Weakly-supervised semantic segmentation network with deep seeded region growing. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Huang et al. 2020] Huang, Y.; Zheng, F.; Cong, R.; Huang, W.; Scott, M. R.; and Shao, L. 2020. Mcmt-gan: Multi-task coherent modality transferable gan for 3d brain image synthesis. *IEEE Transactions on Image Processing* 29:8187–8198.

[Huang et al. 2021] Huang, Y.; Zheng, F.; Wang, D.; Huang, W.; Scott, M. R.; and Shao, L. 2021. Brain image synthesis with unsupervised multivariate canonical cscl4net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5881–5890.

[Hung et al. 2018] Hung, W. C.; Tsai, Y. H.; Liou, Y. T.; Lin, Y. Y.; and Yang, M. H. 2018. Adversarial learning for semi-supervised semantic segmentation.

[Ke et al. 2020] Ke, Z.; Di Qiu, K. L.; Yan, Q.; and Lau, R. W. 2020. Guided collaborative training for pixel-wise semi-supervised learning. In *ECCV*, volume 2, 6. Springer.

[Krizhevsky, Sutskever, and Hinton 2012] Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25:1097–1105.

[Laine and Aila 2016] Laine, S., and Aila, T. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

[Lee et al. 2019] Lee, J.; Kim, E.; Lee, S.; Lee, J.; and Yoon, S. 2019. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference.

[Liu et al. 2021] Liu, S.; Zhi, S.; Johns, E.; and Davison, A. J. 2021. Bootstrapping semantic segmentation with regional contrast. *arXiv preprint arXiv:2104.04465*.

[Long, Shelhamer, and Darrell 2015] Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

[Luo et al. 2020] Luo, X.; Chen, J.; Song, T.; Chen, Y.; Wang, G.; and Zhang, S. 2020. Semi-supervised medical image segmentation through dual-task consistency. *arXiv preprint arXiv:2009.04448*.

[Mittal, Tatarchenko, and Brox 2019] Mittal, S.; Tatarchenko, M.; and Brox, T. 2019. Semi-supervised semantic segmentation with high- and low-level consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP(99):1–1.

[Olsson et al. 2021] Olsson, V.; Tranheden, W.; Pinto, J.; and Svensson, L. 2021. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1369–1378.

[Ouali, Hudelot, and Tami 2020] Ouali, Y.; Hudelot, C.; and Tami, M. 2020. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12674–12684.

[Papandreou et al. 2016] Papandreou, G.; Chen, L. C.; Murphy, K. P.; and Yuille, A. L. 2016. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *IEEE International Conference on Computer Vision*.

[Ronneberger, Fischer, and Brox 2015a] Ronneberger, O.; Fischer, P.; and Brox, T. 2015a. U-net: Convolutional networks for biomedical image segmentation. *ArXiv* abs/1505.04597.

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

[Sajjadi, Javanmardi, and Tasdizen 2016] Sajjadi, M.; Javanmardi, M.; and Tasdizen, T. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning.

[Shi et al. 2016] Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1874–1883.

[Sohn et al. 2020] Sohn, K.; Berthelot, D.; Li, C.-L.; Zhang, Z.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Zhang, H.; and Raffel, C. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*.

[Souly, Spampinato, and Shah 2017] Souly, N.; Spampinato, C.; and Shah, M. 2017. Semi supervised semantic segmentation using generative adversarial network. In *2017 IEEE International Conference on Computer Vision (ICCV)*.

[Takeru et al. 2018] Takeru, M.; Shin-Ichi, M.; Shin, I.; and Masanori, K. 2018. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1–1.

[Tarvainen and Valpola 2017a] Tarvainen, A., and Valpola, H. 2017a. Mean teachers are better role models: Weight-

averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*.

[Tarvainen and Valpola 2017b] Tarvainen, A., and Valpola, H. 2017b. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results.

[Wang et al. 2018] Wang, X.; Girshick, R. B.; Gupta, A.; and He, K. 2018. Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 7794–7803.

[Wei et al. 2018] Wei, Y.; Xiao, H.; Shi, H.; Jie, Z.; Feng, J.; and Huang, T. S. 2018. Revisiting dilated convolution: A simple approach for weakly- and semi- supervised semantic segmentation.

[Zhai et al. 2019] Zhai, X.; Oliver, A.; Kolesnikov, A.; and Beyer, L. 2019. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1476–1485.

[Zhang et al. 2017] Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

[Zhao et al. 2017] Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.