arXiv:2012.04170v1 [cs.CV] 8 Dec 2020

Weakly-Supervised Cross-Domain Adaptation for Endoscopic Lesions Segmentation

Jiahua Dong, Yang Cong, *Senior Member, IEEE,* Gan Sun, Yunsheng Yang, Xiaowei Xu, and Zhengming Ding

***Abstract*—Weakly-supervised learning has attracted growing research attention on medical lesions segmentation due to significant saving in pixel-level annotation cost. However, 1) most existing methods require effective prior and constraints to explore the intrinsic lesions characterization, which only generates incorrect and rough prediction; 2) they neglect the underlying semantic dependencies among weakly-labeled target enteroscopy diseases and fully-annotated source gastroscope lesions, while forcefully utilizing untransferable dependencies leads to the negative performance. To tackle above issues, we propose a new weakly-supervised lesions transfer framework, which can not only explore transferable domain-invariant knowledge across different datasets, but also prevent the negative transfer of untransferable representations. Specifically, a Wasserstein quantified transferability framework is developed to highlight wide-range transferable contextual dependencies, while neglecting the irrelevant semantic characterizations. Moreover, a novel self-supervised pseudo label generator is designed to equally provide confident pseudo pixel labels for both hard-to-transfer and easy-to-transfer target samples. It inhibits the enormous deviation of false pseudo pixel labels under the self-supervision manner. Afterwards, dynamically-searched feature centroids are aligned to narrow category-wise distribution shift. Comprehensive theoretical analysis and experiments show the superiority of our model on the endoscopic dataset and several public datasets.**

***Index Terms*—Weakly-supervised learning, endoscopic lesions segmentation, semantic knowledge transfer, domain adaptation.**

I. Introduction

WEAKLY-SUPERVISED learning has been successfully extended into various computer vision tasks, especially for expensive tasks in terms of pixel-level annotations, *e.g.,* semantic segmentation of medical images [1]–[3]. It has made great contributions to assisting the clinical experts by learning a semantic lesions segmentation model for pixel-level predictions with only accessing weakly-annotated (image-level) labels. To improve the accuracy and efficiency of medical lesions diagnosis, it has been successfully applied into a large amount of clinical diagnosis tasks until now, *e.g.,* thoracic disease localization [4], automated glaucoma detection [5], histopathology segmentation [1], etc.

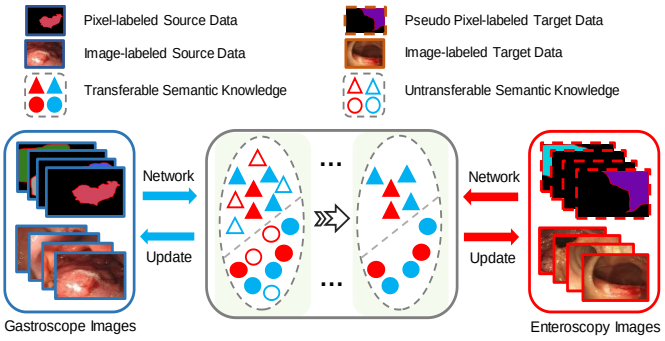


Fig. 1. Illustration of our weakly-supervised semantic knowledge transfer framework. It selectively highlights transferable semantic characterizations from source lesions to assist the target diseases segmentation while neglecting these untransferable representations.

However, when required to learn lesions locations by providing image category supervision, weakly-supervised semantic segmentation of medical lesions is still a thorny diagnosis task since: 1) prior distribution of lesions and clinical expertise [1] are required to directly associate high-level semantic representation with low-level appearance, which cannot produce discriminative and correct lesions locations; 2) it neglects relevant semantic dependencies among target weakly-labeled diseases and source finely-labeled lesions. Obviously, semantic knowledge between two different domains (*e.g.*, enteroscopy and gastroscope datasets) may be not all equally transferable, while forcefully adapting the untransferable dependencies is vulnerable to negative transfer of irrelevant knowledge. Furthermore, it is thorny to manually determine what kinds of semantic dependencies are transferable or not. For example, endoscopic lesions with various appearances and different backgrounds in the images could not contribute equally to semantic transfer in the feature space. Therefore, in this paper, how to establish a weakly-supervised semantic lesions transfer model with highly-selective capacity for transferable knowledge while ignoring irrelevant semantic knowledge among different domains is our main focus.

To tackle above challenges, as depicted in Fig. 1, we

This work is supported by Ministry of Science and Technology of the People´s Republic of China (2019YFB1310300), National Nature Science Foundation of China under Grant (61722311, U1613214, 61821005, 61533015) and National Postdoctoral Innovative Talents Support Program (BX20200353). *(Corresponding author: Yang Cong.)*

Jiahua Dong is with the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China, and also with the Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110016, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: dongjiahua@sia.cn).

Yang Cong and Gan Sun are with the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China, and also with the Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110016, China (e-mail: congyang81@gmail.com, sungan1412@gmail.com).

Yunsheng Yang is with the Chinese PLA General Hospital, Beijing 100000, China (email: sunny301ddc@126.com).

Xiaowei Xu is with the Department of Information Science, University of Arkansas at Little Rock, Arkansas 72204, USA (e-mail: xwxu@ualr.edu).

Zhengming Ding is with the Department of Computer, Information and Technology, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202 USA (email: zd2@iu.edu).

This is the author's manuscript of the article published in final edited form as:

Dong J, Cong Y, Sun G, Yang Y, Xu X, Ding Z. Weakly-Supervised Cross-Domain Adaptation for Endoscopic Lesions Segmentation. IEEE Transactions on Circuits and Systems for Video Technology. 2021;31(5):2020-2033. doi:10.1109/TCSVT.2020.3016058

propose a new weakly-supervised lesions transfer model, whose goal is to exploit transferable semantic knowledge between fully-labeled source gastroscope domain and weakly-labeled target enteroscopy domain, and brush irrelevant knowledge aside in a highly-selective manner. To be specific, a Wasserstein adversarial learning based quantified transferability framework guides our model to focus on important semantic knowledge with high transfer scores, while ignoring these irrelevant dependencies. It incorporates two types of complementary parallel self-attention mechanisms to adaptively capture long-range contextual dependencies, and aligns the heterogeneous probability distributions among different datasets under the Wasserstein distance metric. In contrast to [3], a self-supervised pseudo pixel label generator under curriculum learning strategy could produce confident pseudo labels by taking into account class balance and super-pixel prior [6]. It pays equal attention to both hard-to-transfer and easy-to-transfer classes of target dataset while preventing the enormous deviation of false pseudo labels and the gradual dominance of easy-to-transfer classes along the training process. Afterwards, in order to narrow the category-wise feature distribution shifts, we align the dynamically-searched feature centroids across domains, which are gradually updated with the assistance of previously learned features under the supervision of pseudo labels. In addition, theoretical analysis about how our model explores domain-invariant transferable knowledge across different datasets is elaborated. Empirical comparisons on endoscopic dataset and some representative non-medical datasets illustrate that our model improves the transfer performance significantly when compared with other state-of-the-art approaches.

The key contributions of this paper are presented as follows:

- We propose a new weakly-supervised lesions transfer framework for endoscopic lesions segmentation. To the best of our knowledge, this is an earlier attempt to selectively explore transferable representations while neglecting irrelevant lesions characterizations in the biomedical image processing field.
- We develop a Wasserstein quantified transferability framework based on two types complementary parallel self-attention mechanisms to selectively transfer important knowledge with high transferability while ignoring those untransferable semantic dependencies.
- A self-supervised pseudo label generator is designed to equally generate confident pseudo pixel labels for both hard-to-transfer and easy-to-transfer classes of target dataset, which can not only inhibit gradual dominance of easy-to-transfer classes along the training process, but also further prevent the enormous deviation of false pseudo labels.
- Comprehensive theoretical analysis guarantees that our model could effectively explore domain-invariant knowledge to minimize the domain discrepancy.

This work is an extension version of our previous conference paper [3] and the main improvements include: 1) A Wasserstein quantified transferability framework is developed to highlight important transferable dependencies while neglecting these untransferable semantic knowledge. It mitigates the heterogeneous probability distributions shift among different datasets under the Wasserstein distance metric. 2) Both spatial attention network and channel attention network are exploited to capture long-range context information for endoscopic lesions with various appearances. 3) Different from our conference version [3], a self-supervised pseudo label generator is developed to progressively produce more confident pseudo labels under a self-supervision way, which is determined by cross-entropy loss rather than probability [3]. It further prevents the enormous deviation of false pseudo labels in the fine-tuning process. 4) Theoretical analysis guarantees that our model could effectively bridge the cross-domain distribution shift to learn transferable knowledge.

II. Related Work

This section detailedly introduces some related researches about semantic segmentation of medical lesions and semantic knowledge transfer.

A. Semantic Segmentation of Medical Lesions

In the past few decades, computer aided diagnosis (CAD) [7] has been widely applied to assist clinician due to its significant improvement in the accuracy and efficiency of medical lesions segmentation. Traditional methods [8], [9] rely on clinical experience [10] and domain experts [11] to handcraft local image features to characterize medical lesions. After deep convolutional neural networks [12]–[14] achieve great successes, many advanced methods [15], [16] relying on powerful nets are introduced to attain better performance for semantic segmentation task of medical lesions. They require large-scale finely-annotated pixel labels to train the whole network architectures. In order to efficiently save pixel annotation efforts, weakly-supervised semantic segmentation [17] focuses on exploring the dense pixel-wise labeling from image-level tags. Specifically, [17] proposes a novel loss function to optimize the output probability space with any set of linear constraints. Inspired by [17], weakly-supervised learning for medical lesions segmentation task [1], [2], [4] has attracted growing research interests. Generally, they require effective prior [4] and constraints [1], [5] to explore discriminative lesions representation associated with image-level labels while producing inaccurate and coarse lesions localization. [18] introduces a differentiable penalty for the loss function to avoid expensive Lagrangian dual iterations and proposal generation. Basically, there is still a significant prediction gap between the training models with pixel-level supervision and image-level guidance.

B. Semantic Knowledge Transfer

After [19] introduces generative adversarial network [20] to minimize feature distribution gap across domains, various variants [21], [22] are proposed for classification adaptation task. As pointed out by Zhang et al. [23], the domain adaptation methods for classification cannot be directly applied into the semantic segmentation task. To this end, [24] proposes to

translate the source data to target data for domain adaptation in segmentation task. Several researches [25]–[32] focus on utilizing generative adversarial architecture to transfer domain-invariant knowledge across domains in the feature space. Besides, Tsai *et al.* [33] develop a multi-level contextual adaptation framework in the output space. Different from them, Zou et al. [34], [35] propose a self-training strategy instead of adversarial learning for semantic representation transfer. [36] considers the bounding box as weak supervision to guide cross-domain urban scenes adaptation. [37] translates source dataset into a continuous sequence of intermediate datasets with different style to smoothly bridge domain gap. [38]–[41] propose some novel adaptation losses to narrow the distribution discrepancy or semantic inconsistency across domains. [42] develops a quantified transferability perception mechanism to highlight transferable representations across gastroscope and enteroscopy domains, while neglecting untransferable knowledge. [43] introduces a similarity guided constraint to map features from the same class across domains nearby. [44] proposes an importance sampling mechanism to quantify sample contributions, which effectively prevents the distractions from the unexpected noisy samples. [45] integrates the category-wise matching strategy into a transferable feature learning module and a robust task predictor.

III. The Proposed Model

In this section, we present brief model overview and detailed formulation of our weakly-supervised lesions transfer framework, followed by the theoretical analysis.

A. Overview

As the overall framework demonstrated in Fig. 2, we respectively design two subnets denoted as S_1 and S_2 for classification and segmentation, where classification probability generated by S_1 is used to refine the pixel prediction of subnet S_2 via convolution operation. Denote the source gastroscope samples and target enteroscopy data as $X^s = \{(x_i^s, y_i^{sc}, y_i^{ss})\}_{i=1}^{n_s} \in P_s$ and $X^t = \{(x_j^t, y_j^{tc})\}_{j=1}^{n_t} \in P_t$, respectively, where y_i^{sc} and y_i^{ss} represent the image-level and pixel-level annotations of source image x_i^s, and y_j^{tc} indicates the image tag of target image x_j^t. n_s and n_t represent the numbers of images for each domain. P_s and P_t represent the source and target marginal distributions. The image x_i^s of source dataset X^s is first passed into the whole architecture excluding discriminator D to optimize network parameters. We then utilize the backbone ResNet-50 to extract semantic features for image x_j^t of target dataset X^t. Discriminator D takes feature maps of source dataset X^s and target dataset X^t as input to identify which one is from X^s or X^t, which encourages both X^s and X^t to share closer feature distribution in the high-level semantic space.

Even though generative adversarial learning is employed to mitigate the domain gap between X^s and X^t, it cannot ensure the category-wise feature distributions across domains are mapped nearby. To this end, we focus on aligning category-wise feature centroids to narrow the distribution shift. Unfortunately, feature centroids for target dataset X^t are difficult to compute due to lack of valid pixel annotations. To tackle this challenge, we develop a novel self-supervised pseudo pixel labels generator that incorporates class balance and super-pixel prior. Relying on the generated pseudo pixel labels for target data X^t, we compute semantic centroids for each class by resorting to previously learned features. However, the semantic representations between gastroscope and enteroscopy domains are not all transferable, while forcefully exploiting these untransferable representations could lead to negative performance. Therefore, as depicted in Fig. 2, a Wasserstein adversarial network based quantified transferability framework is designed to explore transferable dependencies in both spatial and channel dimensions, while neglecting the irrelevant representation.

B. Model Formulation

To explore transferable knowledge for target lesions segmentation task in a highly-selective manner, our proposed model is formulated as follows:

$$\begin{aligned}\mathcal{L} =&\mathcal{L}_C(X^s, X^t) + \mathcal{L}_S(X^s, X^t)+\\ &\eta\mathcal{L}_W(P_s, P_t) + \mu\mathcal{L}_T(X^s, X^t),\end{aligned} \tag{1}$$

where $\eta \geq 0$ and $\mu \geq 0$ represent the balanced hyper-parameters. Each loss formulation in Eq. (1) can be defined as follows:

1) Classification Loss $\mathcal{L}_C(X^s, X^t)$: The classification loss for both source gastroscope data and target enteroscopy data is expressed as $\mathcal{L}_C(X^s, X^t)$. The subnet S_1 is trained with loss $\mathcal{L}_C(X^s, X^t)$ to distinguish whether the input image has lesions regions or not. $\mathcal{L}_C(X^s, X^t)$ can be concretely written as:

$$\begin{aligned}\mathcal{L}_C(X^s, X^t) =&\mathbb{E}_{(x_i^s, y_i^{sc})\in X^s}\big(-(y_i^{sc})^\top \log(S_1(x_i^s, \theta_{S_1}))\big)\\ &+\mathbb{E}_{(x_j^t, y_j^{tc})\in X^t}\big(-(y_j^{tc})^\top \log(S_1(x_j^t, \theta_{S_1}))\big),\end{aligned} \tag{2}$$

where θ_{S_1} indicates the network weights of the classifier S_1. $S_1(x_i^s, \theta_{S_1})$ and $S_1(x_j^t, \theta_{S_1})$ respectively denote the corresponding softmax outputs of classification network for source and target images.

2) Segmentation Loss $\mathcal{L}_S(X^s, X^t)$: Given the predicted softmax outputs of the subnet S_2, we formulate $\mathcal{L}_S(X^s, X^t)$ as the segmentation loss for both source dataset X^s with finely-annotated pixel-level labels y_i^{ss}, and target dataset X^t with generated highly-confident pseudo pixel annotations. $\mathcal{L}_S(X^s, X^t)$ can then be expressed as:

$$\begin{aligned}\mathcal{L}_S(X^s, X^t) =& \mathbb{E}_{(x_i^s, y_i^{ss})\in X^s}\big(-\sum_{a=1}^{|x_i^s|}(y_{ia}^{ss})^\top \log(S_2(x_i^s, \theta_{S_2})_a)\big)\\ &+\mathbb{E}_{x_j^t\in X^t}\big(-\sum_{b=1}^{|x_j^t|} v_j^b(\hat{y}_{jb}^{ts})^\top \log(S_2(x_j^t, \theta_{S_2})_b) - \lambda v_j^b\big),\\ s.t.,\ &\hat{y}_{jb}^{ts} \in \big\{\mathbf{e}_k | \mathbf{e}_k \in \mathbb{R}^K, k = \arg\max_k S_2(x_j^t, \theta_{S_2})_b\big\},\\ &v_j^b \in \{0, 1\},\ \forall b = 1, \ldots |x_j^t|,\end{aligned} \tag{3}$$

where θ_{S_2} represents the network parameters of S_2. $S_2(x_i^s, \theta_{S_2})_a$ and $S_2(x_j^t, \theta_{S_2})_b$ respectively denote the classes

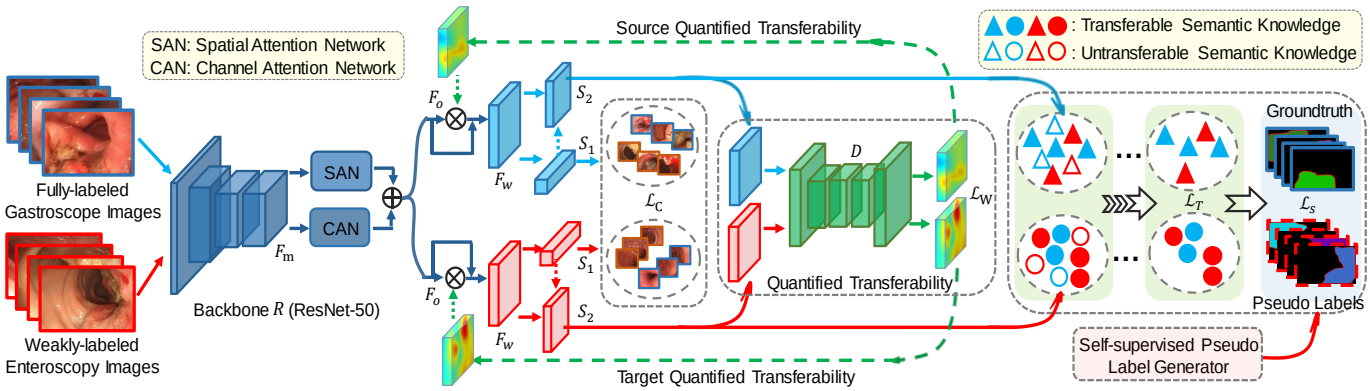


Fig. 2. Overview of our proposed model that mainly contains three significant modules, *i.e.*, *a Wasserstein adversarial learning based quantified transferability* $\mathcal{L}_W$ for exploring transferable lesions dependencies while ignoring irrelevant knowledge, *a self-supervised pseudo label generator* to produce confident pseudo labels for weakly-labeled target samples, and *a semantic consistent loss* $\mathcal{L}_T$ to align category-wise feature centroids across domains.

probabilities predicted by pixel-wise classifier S_2 at pixel a ($a = 1, 2, ..., |x_i^s|$) and b ($b = 1, 2, ..., |x_j^t|$). y_{ia}^{ss} and $\hat{y}_{jb}^{ts}$ are the one-hot groundtruth and confident pseudo label for the a-th and the b-th pixel positions in source image x_i^s and target sample x_j^t, respectively. K and $\mathbf{e}_k$ indicate the number of categories and one-hot vector. v_j^b is the weight of the b-th pixel in target sample x_j^t. Notice that assigning v_j^b as 0 can ignore its influence on training process. Thus, we expect that $\lambda \geq 0$ can serve as a global trade-off controller to determine the selection scale of pseudo labels for target dataset, and prevent the trivial solution from assigning all generated pseudo labels as 0. Obviously, a larger λ can promote the selection scale of confident pseudo pixel labels.

Although Eq. (3) could progressively mine pseudo pixel labels with high confidence score along the optimization process, the selection manner of pseudo labels in Eq. (3) still need to overcome three drawbacks: **(i)** Eq. (3) will have a preference for initially easier-to-learn classes while ignoring those hard-to-transfer categories in initial optimization procedure; **(ii)** The false pseudo labels could heavily cripple the performance along the training process; **(iii)** The selected pseudo pixel labels used for training are spatially discrete.

In order to cope with the issues **(i)** and **(ii)**, the second optimization term in Eq. (3) could be concretely rewritten as follows:

$$\begin{aligned}
&\min_{v_b} \mathbb{E}_{x_j^t \in X^t}\Big(-\sum_{b=1}^{|x_j^t|}\sum_{k=1}^{K} p_j^b \left[v_j^b (\hat{y}_{jb}^{ts})_k^\top \log(S_2(x_j^t, \theta_{S_2})_b)\right] - \lambda_k v_j^b\Big), \\
&s.t.,\ \ \hat{y}_{jb}^{ts} \in \left\{\mathbf{e}_k | \mathbf{e}_k \in \mathbb{R}^K, k = \arg\max_k S_2(x_j^t, \theta_{S_2})_b\right\}, \\
&p_j^b = \max_k S_2(x_j^t, \theta_{S_2})_b, v_j^b \in \{0, 1\},\ \ \forall b = 1, \ldots |x_j^t|,
\end{aligned} \tag{4}$$

where λ_k ($k = 1, 2, ..., K$) are class-balanced weights controlling selection scale of produced pseudo labels with respect to each class k. The cross-entropy loss of the b-th pixel is weighted by its maximum class probability p_j^b, which further averts the enormous deviation of false pseudo labels when retraining the whole network in a self-supervised manner. To prevent the dominance of large number of pixel categories, different from [3], cross-entropy loss of each pixel is employed to determinate the value of λ_k, as summarized in **Algorithm 1**. After computing cross-entropy loss L_j for each pixel of all target images, we sort the losses SL_k weighted by the predicted maximal probability of corresponding pixel for each class k. Then λ_k can be determined by the self-supervised cross-entropy loss ranked at $n\text{length}(SL_k)$. Besides, the value of n starts from 25% and increases by 5% in each iterative optimization epoch, and we set the maximum value of n as 55%. The intuitive explanation about iteratively adding the value of n is that the trained model will generate more confident pseudo labels as the training process. Moreover, the optimal solution of Eq. (4) can be expressed as:

$$v_j^b = \begin{cases} 1, & \text{if } \ (\hat{y}_{jb}^{ts})^\top \log(S_2(x_j^t, \theta_{S_2})_b) > -\dfrac{\lambda_k}{p_j^b}, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

To tackle the issue **(iii)**, the weight parameter $v_j = \{v_j^b\}_{b=1}^{|x_j^t|}$ of target image x_j^t that are determined by Eq. (5) can be further modified by super-pixel prior [6]. It encourages spatial contextual continuity of selected pseudo labels for model retraining. Furthermore, the procedure about how to refine the weight parameter v_j via super-pixel prior is elaborated in **Algorithm 2**, where S_j^t represents super-pixel prior of target image x_j^t. When the weight parameter v_j at the (h, w)-th pixel in target image x_j^t equals to 0, its ultimate refinement can be determined by the information of its 8-neighborhoods under the voting strategy.

3) Wasserstein Adversarial Loss $\mathcal{L}_W(P_s, P_t)$: Due to the heterogeneous probability distribution of different datasets, Wasserstein-1 distance [46] is employed to better measure the distribution discrepancy across P_s and P_t when comparing with other distance functions. The Wasserstein-1 discrepancy $\mathcal{L}_W(P_s, P_t)$ between P_s and P_t can be defined as follows:

$$\mathcal{L}_W(P_s, P_t) = \inf \mathbb{E}_{(x_i^s, x_j^t) \in \Pi(P_s, P_t)}(|S_2(x_i^s, \theta_{S_2}) - S_2(x_j^t, \theta_{S_2})|), \tag{6}$$

where $\Pi(P_s, P_t)$ denotes the joint distribution whose marginal distributions are P_s and P_t. $S_2(x_i^s, \theta_{S_2})$ and $S_2(x_j^t, \theta_{S_2})$ respectively represent the extracted features for source and target samples. θ_{S_2} is the network weight of subnet S_2. Since

Algorithm 1 Determination Strategy of λ_k

Input: The generation portion n of pseudo labels, target enteroscopy sample $x_j^t \in X^t$, K, S_2;
Output: λ_k

for $j = 1, \ldots, |X^t|$ **do**
  $Y_j = \text{argmax}(S_2(x_j^t, \theta_{S_2}), \text{axis} = 3)$;
  $M_j = \max(S_2(x_j^t, \theta_{S_2}),\ \text{axis} = 3)$;
  $L_j = \text{pixel_cross_entropy_loss}(Y_j, S_2(x_j^t, \theta_{S_2}))$;
  $ML_j = M_j \circ L_j$, where $\circ$ denotes Hadamard product;
  for $k = 1, \ldots, K$ **do**
    $ML_j^k = ML_j(Y_j == k)$;
    $PL_k = [PL_k, \text{vector}(ML_j^k)]$;
  end for
end for
for $k = 1, \ldots, K$ **do**
  $SL_k = \text{ascending_sorting}(PL_k)$;
  $T_k = n\text{length}(SL_k)$;
  $\lambda_k = SL_k[T_k]$;
end for
return λ_k;

Algorithm 2 Selection Strategy of Weight Parameter v_j

Input: Height I_h and width I_w of target enteroscopy sample $x_j^t \in X^t$, K;
Output: Weight parameter v_j;

Determine λ_k via **Algorithm 1**;
for $j = 1, \ldots, |X^t|$ **do**
  Solve initial weight parameter v_j of x_j^t via Eq. (5);
  Solve super-pixel prior S_j^t of x_j^t via [6];
  for $h = 1, \ldots, I_h,\ w = 1, \ldots, I_w$ **do**
    Set $C_{hw} = \emptyset$;
    if v_j equals to 0 at (h, w)-th pixel of x_j^t **then**
      for $k = 1, \ldots, K$ **do**
      $A = \{h-1, h, h+1\}, B = \{w-1, w, w+1\}$;
      $C_{hw}^k = \sum_{x \in A} \sum_{y \in B} \mathbf{1}_{((\hat{y}_j^{ts})_{xy}=k)\&((S_j^t)_{hw}=(S_j^t)_{xy})\&((v_j)_{xy}=1)}$;
      $C_{hw} = [C_{hw},\ C_{hw}^k]$;
      end for
      $N_k = \text{argmax}(C_{hw},\ \text{axis} = 0)$;
      if $C_{hw}[N_k] > 4$ **then**
        $(v_j)_{hw} = 1$;
      end if
    end if
  end for
  Return the ultimate weight parameter v_j for x_j^t;
end for

the infimum of Eq. (6) is difficult to solve directly, $\mathcal{L}_W(P_s, P_t)$ in Eq. (6) is equally reformulated as below:

$$\sup_{||D \leq 1||} \left| \mathbb{E}_{x_i^s \in P_s}\big(D(S_2(x_i^s, \theta_{S_2}))\big) - \mathbb{E}_{x_j^t \in P_t}\big(D(S_2(x_j^t, \theta_{S_2}))\big) \right|, \tag{7}$$

where D represents the domain critic function that takes the source and target features as input to identify which one is from X^s or X^t. In this paper, we utilize D as discriminator with the network parameter as θ_D. Consequently, Eq. (7) can be approximately rewritten as:

$$\begin{aligned} &\min_{\theta_{S_2}} \max_{\theta_D} \mathcal{L}_W(P_s, P_t) = \\ &\frac{1}{n_s} \sum_{x_i^s \in X^s} D(S_2(x_i^s, \theta_{S_2}), \theta_D) - \frac{1}{n_t} \sum_{x_j^t \in X^t} D(S_2(x_j^t, \theta_{S_2}), \theta_D) \\ &- \xi\big((||\nabla_{\theta_D} D(x_i^s, \theta_D)|| - 1)^2 - (||\nabla_{\theta_D} D(x_j^t, \theta_D)|| - 1)^2\big), \end{aligned} \tag{8}$$

where $\xi \geq 0$ is the balanced weight.

However, the typical Wasserstein adversarial objective (*i.e.*, Eq. (8)) cannot highlight transferable knowledge while ignoring the untransferable dependencies. Therefore, we develop a Wasserstein adversarial network based quantified transferability mechanism to selectively explore transferable semantic representation between X^s and X^t. Specifically, as shown in Fig. 2, two types of complementary parallel self-attention networks (SAN and CAN) are integrated into generator (*i.e.*, ResNet-50) to capture widely-separated contextual dependencies in both spatial and channel dimensions. Besides, discriminator D is used to adaptively highlight long-range transferable representations and neglects irrelevant knowledge according to its output scores. The details about learning transferable dependencies for both generator and discriminator are shown as follows:

Generator Backbone R**:** As illustrated in Fig. 2, we integrate both Spatial Attention Network (SAN) and Channel Attention Network (CAN) into the ResNet-50 [12], where the details about SAN and CAN are depicted in Fig. 3.

For spatial attention network (SAN) demonstrated in Fig. 3 (a), given the intermediate feature $F_m \in \mathbb{R}^{H \times W \times C}$ extracted by ResNet-50 [12], we respectively feed it into two convolution layers to get two new features $C_1 \in \mathbb{R}^{H \times W \times C}$ and $C_2 \in \mathbb{R}^{H \times W \times C}$, where H, W and C indicate the height, width, channels of F_m, C_1 and C_2. After reshaping C_1 and C_2 into $\mathbb{R}^{N \times C}$ where $N = H \times W$ represents the total number of feature positions from hidden convolutional layer, we then apply a multiplication operation between C_1 and the transpose of C_2. Afterwards, softmax operator is utilized to calculate the position attention matrix $M = \{M_{ij}\}_{i,j=1}^N \in \mathbb{R}^{N \times N}$ and the (i, j)-th item of M is formulated as: $M_{ij} = \frac{\exp\left((C_1 C_2^\top)_{ij}\right)}{\sum_{i=1}^N \exp\left((C_1 C_2^\top)_{ij}\right)}$, where M_{ij} measures the semantic representation correlation between different positions of corresponding feature map. The greater semantic dependency among them contributes to the larger value of M_{ij}.

Furthermore, feature F_m is also forwarded into another convolutional block to produce a new feature $C_3 \in \mathbb{R}^{H \times W \times C}$. We reshape it to $\mathbb{R}^{N \times C}$, and perform a multiplication operation among reshaped C_3 and the transpose of M to output the attention feature map. Afterwards, we multiply the reshaped attention feature map ($\in \mathbb{R}^{H \times W \times C}$) by a scalar weight δ, and conduct an element-wise sum operation with F_m to output $F_p \in \mathbb{R}^{H \times W \times C}$:

$$(F_p)_i = \delta \sum_{j=1}^{N} M_{ij} (C_3)_j + (F_m)_i, \tag{9}$$

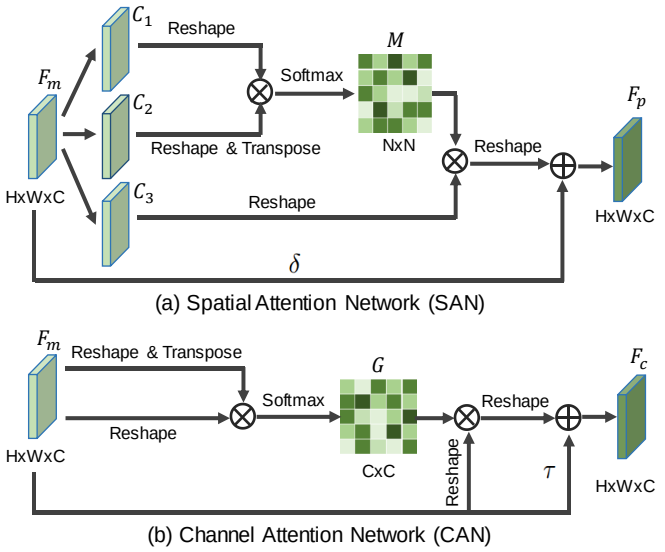


Fig. 3. Detailed demonstration of (a) Spatial Attention Network (SAN) and (b) Channel Attention Network (CAN) in Fig. 2.

where $(F_p)_i, (F_m)_i$ and $(C_3)_j$ respectively represent the semantic representations at the i-th and j-th locations in the corresponding feature maps. Motivated by [47], we initialize δ as 0 and its value is learned adaptively.

Furthermore, for channel attention network (CAN) depicted in Fig. 3 (b), channel attention $G \in \mathbb{R}^{C\times C}$ can be directly calculated with the input feature map $F_m \in \mathbb{R}^{H\times W\times C}$. Concretely, after reshaping feature F_m to $\mathbb{R}^{N\times C}$, we employ a multiplication operator between the transpose of F_m and F_m, followed by softmax operation to calculate the channel attention $G = \{G_{ij}\}_{i,j=1}^{C} \in \mathbb{R}^{C\times C}$, where $G_{ij} = \frac{\exp\left((F_m^\top F_m)_{ij}\right)}{\sum_{i=1}^{C}\exp\left((F_m^\top F_m)_{ij}\right)}$ denotes interdependent correlation between different channels of feature map F_m. Besides, we utilize the matrix multiplication between F_m and the transpose of channel attention G, and their result is then reshaped to $\mathbb{R}^{H\times W\times C}$. Afterwards, we multiply their reshaped matrix by a scale weight τ, and utilize an element-wise sum operation with the original input F_m to obtain ultimate feature map $F_c \in \mathbb{R}^{H\times W\times C}$:

$$(F_c)_i = \tau \sum_{j=1}^{C} G_{ij}(F_m)_j + (F_m)_i, \tag{10}$$

where $(F_c)_i$ and $(F_m)_j$ represent the semantic representations at the i-th and the j-th locations in the feature maps F_c and F_m, respectively. Similar to δ in Eq. (9), τ is an adaptive parameter and initialized as 0.

Discriminator D: Intuitively, the discriminative ability of domain discriminator D could evaluate whether the representations across domains are transferable or untransferable. Take an example for illustration, the input features from different domains which are already aligned will confuse D for discriminating which one is from X^s or X^t. That is to say, the output probabilities of discriminator D for domain recognition help us to easily distinguish whether the input feature map could be transferred or not. Consequently, to focus on those transferable semantic characterizations, uncertainty evaluation strategy (*i.e*, entropy criterion $H(p) = -\sum_m p_m \log(p_m)$) is used to identify the transferability of each semantic representation. Given the output probability $D(S_2(x_i^s, \theta_{S_2}), \theta_D)$ of discriminator D, the quantified transferability for the source data x_i^s could concretely expressed as:

$$W = 1 - H(D(S_2(x_i^s, \theta_{S_2}), \theta_D)). \tag{11}$$

Similarly, the quantified transferability mechanism is also suitable for the target sample x_j^t, as shown in Fig. 2. However, incorrect transferability quantification may degrade the transfer performance significantly, so we develop a residual transferability quantification mechanism to mitigate such negative effect, *i.e.*,

$$F_w = (1 + W) \circ F_o, \tag{12}$$

where $\circ$ denotes Hadamard product. F_w is shared feature map with quantified transferability, which captures transferable semantic representations, as presented in Fig. 2.

4) Semantic Consistence Loss $\mathcal{L}_T(X^s, X^t)$: To mitigate the feature distribution shift of same class across different datasets X^s and X^t, we present $\mathcal{L}_T(X^s, X^t)$ to achieve semantic knowledge transfer, which aligns the feature centroids with previously-learned representations:

$$\mathcal{L}_T(X^s, X^t) = \sum_{k=1}^{K} \left\|C_k^s - C_k^t\right\|_2^2 + \alpha \left\|C_k^s - C_k^t\right\|_1, \tag{13}$$

where C_k^s and C_k^t respectively denote the feature centroids of the k-th class in source and target domains. $\alpha \geq 0$ aims to balance the sparse feature property. Similarly to the exponential reward strategy in reinforcement learning [48], we develop the **Algorithm 3** to compute feature centroids of each class with the assistance of exponentially-weighted previously learned features in each iteration. Meanwhile, the produced pseudo pixel labels via **Algorithm 2** are employed to conduct feature centroids alignment process.

The intuitive explanation for resorting previously-learned experience rather than directly aligning those newly calculated centroids is straightforward: 1) Category information among each mini-batch is not usually sufficient, *e.g.*, some categories may be missing due to the random selection strategy; 2) The deviation of false pseudo pixel labels will greatly increase when the mini-batch size is small.

C. Details of Network Architecture

1) Backbone R, Subnet S_1 and S_2: Following the framework presented in [3], we employ DeepLab-v3 [49] with ResNet-50 network [12] as our baseline architecture. To output shared higher dimensional feature map for S_1 and S_2, we set the strides of the last two convolution layers of ResNet-50 as 1, and remove the last classification head. Besides, in the last convolution layer of ResNet-50, we adopt 3 dilated convolution kernel with the stride as $\{1, 2, 4\}$ to expand receptive field. The shared semantic feature produced by backbone R is then respectively forwarded into two parallel attention networks (*i.e.*, SAN and CAN), which explores wide-range contextual dependencies via an element-wise feature

Algorithm 3 Optimize Semantic Consistence Loss $\mathcal{L}_T$

Input: $\{C_k^s\}_{k=1}^K, \{C_k^t\}_{k=1}^K$, MAX-ITER;
Output: $\mathcal{L}_T(X^s, X^t)$;
for $t = 1, 2, \ldots,$ MAX-ITER **do**
  $\mathcal{L}_T(X^s, X^t) = 0$;
  $\big((x_i^s, y_i^{ss}), (x_j^t)\big) = \text{randomly_sampling}(X^s, X^t)$;
  $v_j = \text{pseudo_labeling}(x_j^t)$ via **Algorithm 2**;
  Extract pixel features F_i^s and F_j^t via S_2;
  for $k = 1, \ldots, K$ **do**
    $C_k^{sn} = \frac{1}{|x_i^s|} \sum_{a=1}^{|x_i^s|} (F_i^s)_a \mathbf{1}_{(y_i^{ss})_a = k}$;
    $C_k^{tn} = \frac{1}{|x_j^t|} \sum_{b=1}^{|x_j^t|} (F_j^t)_b \mathbf{1}_{(\hat{y}_j^{ts})_b = k}$;
    $C_k^s = \sum_{x=1}^{n} C_k^{sx} \cdot \gamma^{t-x}$; (Exponentially-weighted)
    $C_k^t = \sum_{x=1}^{n} C_k^{tx} \cdot \gamma^{t-x}$; (Exponentially-weighted)
  end for
  Return $\mathcal{L}_T(X^s, X^t)$;
end for

fusion, as depicted in Fig. 2. Afterwards, we forward the fused feature into subnet S_1 for classification task, and highlight transferable dependencies via the transferability quantified by discriminator D before passing it into subnet S_2 for pixel-level prediction. S_2 is composed of an Atrous Spatial Pyramid Pooling (ASPP) module [50] and a pixel-level predictor.

2) Discriminator D: Motivated by [51], for the main components of discriminator D, five fully convolution networks with kernel size as 3 are used for retaining global information when compared with fully connected network. Specifically, the channel numbers of 5 convolutional filters are {16, 32, 64, 64, 1}, respectively. We set the stride of first three convolutional filters as 2 and the stride of other layers as 1. The Leaky RELU function with the parameter as 0.2 is applied to activate each convolution block excluding the last one that is activated by sigmoid function.

D. Training and Testing

1) Training Phase: The overall objective of training our model is formulated as:

$$\min_{\theta_R, \theta_{S_1}, \theta_{S_2}} \max_{\theta_D} \quad \mathcal{L}_{obj} = \mathcal{L}_C + \mathcal{L}_S + \eta \mathcal{L}_W + \mu \mathcal{L}_T. \tag{14}$$

In each iteration step, we use ResNet-50 to extract feature maps of source (*e.g.*, gastroscope) and target (*e.g.*, enteroscopy) images, and then forward them into discriminator D for optimizing $\mathcal{L}_W(P_s, P_t)$. To optimize classification loss $\mathcal{L}_C(X^s, X^t)$ and segmentation loss $\mathcal{L}_S(X^s, X^t)$, source sample x_i^s with the transferability quantified by D is passed into the subnets S_1 and S_2. Afterwards, we obtain the target softmax output $S_2(x_j^t, \theta_{S_2})$ for image x_j^t only with the image-level annotation y_j^{tc}, and corresponding ultimate weight parameter v_j is produced via **Algorithm 2**. Additionally, with assistance of previously learned features, the category feature centroids C_k^s and C_k^t are solved via **Algorithm 3** to optimize the loss $\mathcal{L}_T(X^s, X^t)$. As for detailed implementation, a Titan V GPU with 12 GB memory is utilized for training and the batch size is set as 4. We employ Adam as the optimizer with initial learning rate as 1.0×10^{-4} to train the whole networks architecture. The rate and step size of Adam optimizer are respectively set as 0.7 and 950.

2) Testing Phase: In the testing stage, we directly forward the target enteroscopy sample x_j^t into the backbone network ResNet-50 followed by the subnets S_1 and S_2 for performance prediction.

E. Theoretical Insights

This subsection introduces comprehensive theoretical analysis about how our model could efficiently narrow the domain discrepancy between source distribution P_s and target distribution P_t. According to [52], given the hypothesis class $\mathcal{H}$, the target expected error $\epsilon_T(h)$ of any classifier $h \in \mathcal{H}$ predicting on target domain can be theoretically bounded by the following formulation:

$$\forall h \in \mathcal{H},\ \epsilon_T(h) \leq \epsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\triangle\mathcal{H}(P_s, P_t)} + \Lambda, \tag{15}$$

where $\epsilon_S(h)$ represents the source expected error of $h \in \mathcal{H}$ predicting on source domain, which could be easily minimized to 0 under the fully-supervised learning. Λ denotes the independent constant. Thus, our model focuses on how to minimize the domain discrepancy $d_{\mathcal{H}\triangle\mathcal{H}(P_s, P_t)}$ between source images $x_i^s \in X^s$ and target samples $x_j^t \in X^t$, which can be concretely defined as:

$$\begin{aligned} d_{\mathcal{H}\triangle\mathcal{H}(P_s,P_t)} &= 2 \sup_{h,h' \in \mathcal{H}} |P_{s(h(x_i^s) \neq h'(x_i^s))} - P_{t(h(x_j^t) \neq h'(x_j^t))}| \\ &= 2 \sup_{h,h' \in \mathcal{H}} |\mathbb{E}_{x_i^s \in X^s}(\mathbf{1}_{(h(x_i^s) \neq h'(x_i^s))}) - \mathbb{E}_{x_j^t \in X^t}(\mathbf{1}_{(h(x_j^t) \neq h'(x_j^t))})| \\ &= 2 \max_{h,h' \in \mathcal{H}} |\mathbb{E}_{x_i^s \in X^s}(\mathbf{1}_{(h(x_i^s) \neq h'(x_i^s))}) - \mathbb{E}_{x_j^t \in X^t}(\mathbf{1}_{(h(x_j^t) \neq h'(x_j^t))})|. \end{aligned} \tag{16}$$

The classifiers $h, h' \in \mathcal{H}$ $(h \neq h')$ in Eq. (16) can be decomposed as the subnets $R \circ S_1$ and $R \circ S_2$. They perform on the high-level semantic features extracted from backbone network R for classification and segmentation, respectively. With the fixed parameters of R, S_1 and S_2, Eq. (16) can then be rewritten as Eq. (17) by employing discriminator D to distinguish different distributions, *i.e.*,

$$\begin{aligned} d_{\mathcal{H}\triangle\mathcal{H}(P_s,P_t)} = 2 \max_{\theta_D} \big\{ &|\mathbb{E}_{x_i^s \in X^s}(\mathbf{1}_{(R\circ S_1(x_i^s) \neq R \circ S_2(x_i^s))}) \\ &- \mathbb{E}_{x_j^t \in X^t}(\mathbf{1}_{(R\circ S_1(x_j^t) \neq R \circ S_2(x_j^t))})| \big\}. \end{aligned} \tag{17}$$

Eq. (17) is in accordance with our formulation in Eq. (8), *i.e.*,

$$\begin{aligned} &\max_{\theta_D} \mathcal{L}_W(P_s, P_t) = \\ &\frac{1}{n_s} \sum_{x_i^s \in X^s} D(S_2(x_i^s, \theta_{S_2}), \theta_D) - \frac{1}{n_t} \sum_{x_j^t \in X^t} D(S_2(x_j^t, \theta_{S_2}), \theta_D) \\ &- \xi\big((||\nabla_{\theta_D} D(x_i^s, \theta_D)|| - 1)^2 - (||\nabla_{\theta_D} D(x_j^t, \theta_D)|| - 1)^2\big). \end{aligned} \tag{18}$$

Considering the goal of minimizing the domain discrepancy $d_{\mathcal{H}\triangle\mathcal{H}(P_s,P_t)}$, we aim to mitigate the domain-wise distribution shift by optimizing R, S_1 and S_2 with the losses $\mathcal{L}_C(X^s, X^t), \mathcal{L}_S(X^s, X^t)$ and $\mathcal{L}_T(X^s, X^t)$, *i.e.*,

$$\begin{aligned} \min_{\theta_R, \theta_{S_1}, \theta_{S_2}} \max_{\theta_D} \quad &\mathcal{L}_C(X^s, X^t) + \mathcal{L}_S(X^s, X^t) \\ &+ \eta \mathcal{L}_W(P_s, P_t) + \mu \mathcal{L}_T(X^s, X^t). \end{aligned} \tag{19}$$

TABLE I
COMPARISON EXPERIMENTS BETWEEN OUR MODEL AND SEVERAL STATE-OF-THE-ART APPROACHES ON MEDICAL ENDOSCOPIC DATASET.

Metrics	BL [49]	CDWS [1]	NMD [26]	Wild [25]	DFN [53]	LtA [33]	CGAN [31]	CLAN [40]	ADV [39]	CRST-LR [35]	CRST-MR [35]	SWE [3]	Ours
$\mathrm{IoU}_n(\%)$	75.13	25.11	81.10	81.58	81.33	81.73	80.32	82.31	83.46	83.98	84.21	84.76	**85.21**
$\mathrm{IoU}_d(\%)$	33.24	15.51	36.85	38.59	37.50	41.10	41.33	42.04	42.63	42.85	42.94	43.16	**43.65**
mIoU(%)	54.19	20.31	58.97	60.09	59.41	61.42	60.82	62.18	63.05	63.42	63.57	63.96	**64.43**

TABLE II
NOTATIONS FOR DIFFERENT MODULES IN OUR MODEL AND EVALUATION.

Notations	Interpretation
BL	Baseline architecture DeepLab-v3 [49] without knowledge transfer
PL	Pseudo labels generator
AL	Wasserstein adversarial learning framework
SRT	Semantic consistence loss $\mathcal{L}_T(X^s, X^t)$ in Eq. (13)
QT	Quantified transferability in Eq. (11)
Ours-woSP	Training our model without the super-pixel prior
mIoU	mean intersection over union (IoU)
IoU_d	intersection over union (IoU) of disease
IoU_n	intersection over union (IoU) of normal

Notice that Eq. (19) well matches our formulation in Section III-B and the training objective Eq. (14) in Section III-D. Therefore, by iteratively updating the parameters of R, S_1, S_2 and D, our model could efficiently narrow the domain discrepancy $d_{\mathcal{H}\triangle\mathcal{H}(P_s,P_t)}$ and achieve the tighter upper bound for $\epsilon_T(h)$.

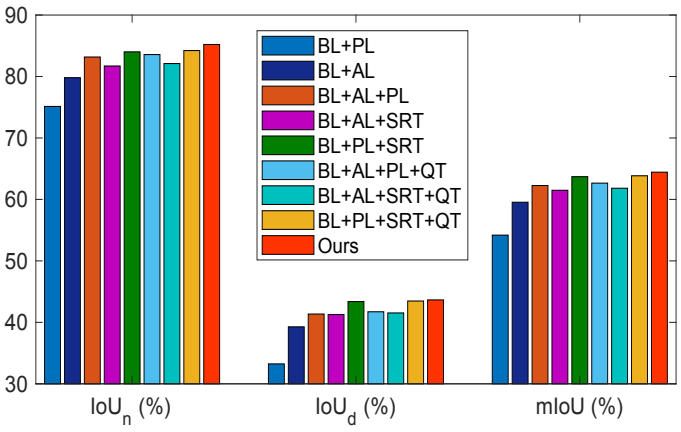


Fig. 4. Ablation studies on medical endoscopic dataset in terms of $\mathrm{IoU}_n, \mathrm{IoU}_d$ and mIoU metrics.

IV. EXPERIMENTS

This section first introduces several representative experimental datasets, and then reports comparison experiments with some state-of-the-art methods to justify the superiority of our model.

A. Dataset and Notations

Medical Endoscopic Dataset [3] is collected from more than 1100 volunteers, and contains 3659 images with various diseases, including cancer, polyp, gastritis, ulcer and bleeding. To be specific, it is composed of 2969 gasteroscope samples with pixel-level annotations and 690 enteroscopy samples only with image-level labels. All gastroscope data and 300 enteroscopy images are respectively regarded as source and target domains for model training. Besides, the rest of enteroscopy samples are employed for evaluation.

Cityscapes [54] with 34 different scene categories is collected from fifty European cities. For training, validation and testing, there are three corresponding subsets with 2993, 1531, 503 samples, respectively.

GTA [55] is composed of 24996 samples that are gathered from the computer game called Grand Theft Auto V. The annotation classes for Cityscapes [54] and **GTA** [55] are compatible with each other.

SYNTHIA [56] is a large-scale synthetic dataset from virtual urban scenes, whose subset named SYNTHIA-RANDCITYSCAPES with total 9400 samples and 12 annotated object categories is employed for performance comparison.

Notations: For simplification, some notations used in this paper are summarized in Table II. To be specific, different components in our model, *i.e.*, baseline architecture DeepLab-v3 [49], pseudo labels generator, Wasserstein adversarial learning, semantic representation transfer, quantified transferability and training without super-pixel prior are abbreviated as BL, PL, AL, SRT, QT and Ours-woSP, respectively. Furthermore, we consider intersection over union (IoU) for basic metric evaluation, where mIoU, IoU_n and IoU_d respectively denote the mean IoU , IoU of normal and IoU of disease.

B. Experiments on Medical Endoscopic Dataset

Several state-of-the-art approaches are employed for experiment comparison, which are presented as follows:

- **Baseline** (**BL**) employs DeepLab-v3 [49] as backbone for segmentation without knowledge transfer.
- **CDWS** [1] explores weakly-supervised multi-scale model for pixel predictions by applying area constraint.
- **NMD** [26] integrates static object prior and soft pseudo labels into multiple class-wise adaptation.
- **Wild** [25] exploits intermediate convolutional blocks by utilizing a adversarial loss based on prior constraint.
- **DFN** [53] focuses on exploring discriminative semantic representations under the supervision manner.
- **LtA** [33] develops a multi-level transfer model in the output space.
- **CGAN** [31] applies conditional GAN into feature space adaptation by incorporating conditional generator.
- **CLAN** [40] exploits category-level alignment network.
- **ADV** [39] proposes entropy minimization objectives based on adversarial loss to narrow the domain gap.
- **CRST** [35] focuses on relaxing the feasible space of pseudo-labels from one-hot vectors to continuous latent variables. We employ the label entropy regularizer (*i.e.*,

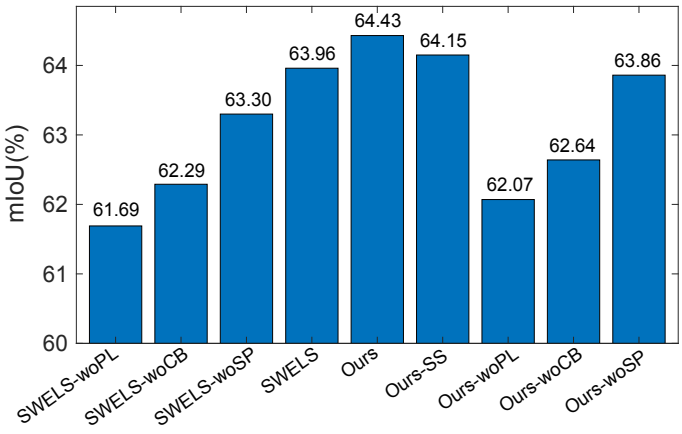


Fig. 5. Discussion of different pseudo label designs on medical dataset.

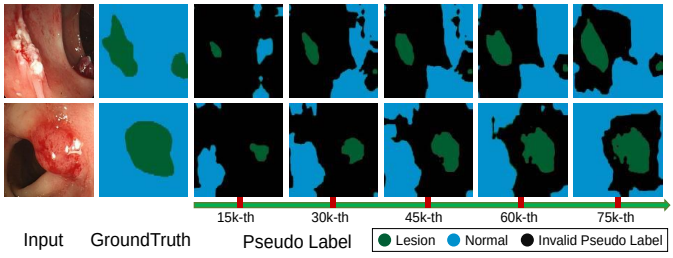


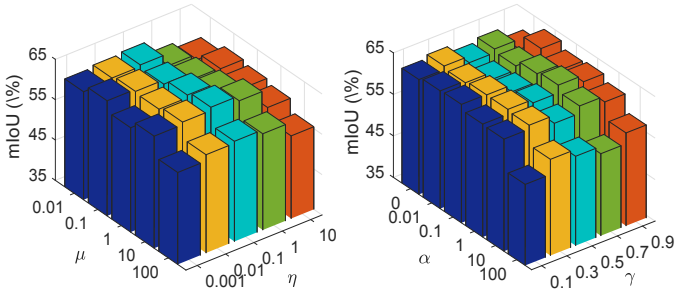
Fig. 6. Examples of intuitively illustrating the propagation process of pseudo labels, where the inputs are enteroscopy samples.

CRST-LR) and model regularizer L_2 (*i.e.*, CRST-MR) for comparison.

- **SWE** [3] is our conference version which develops a confident pseudo label generator to pick highly-confident samples into training set.

For performance comparison, we utilize DeepLab-v3 [49] with ResNet-50 [12] as baseline network, and add an auxiliary image classifier to assist pixel predictions in this experiment. Table I reports the experimental results of our model against several competing approaches. Some observations from Table I are presented as follows: 1) When compared with our conference version [3], the quantified transferability of our model plays an essential role in highlighting wide-range transferable dependencies for both normal and disease classes, even though the medical lesions across different datasets have different appearances and backgrounds. 2) Compared with [3], pseudo label generator can further prevent the enormous deviation of false pseudo labels in the fine-tuning phase when trained under the self-supervision manner. 3) Our model outperforms the state-of-the-arts [31], [33], [39], [40] by a large margin around 1.38∼3.61% in terms of mIoU since the self-supervised pseudo label generator could mine confident pseudo labels for target samples progressively. 4) Models with semantic transfer [25], [26], [31], [33], [39], [40], [53] perform better than those without semantic transfer [1], [49].

1) Ablation Study: This subsection investigates the effectiveness of different components of our proposed model. Fig. 4 reports the performance in term of mIoU achieved by different variants settings on medical dataset. From the



(a) when $\gamma = 0.7, \alpha = 1$ (b) when $\mu = 10, \eta = 0.3$

Fig. 7. The parameter investigations with respect to $\{\mu, \eta\}$ (a) and $\{\alpha, \gamma\}$ (b) on medical endoscopic dataset.

TABLE III
THE EFFECT OF QUANTIFIED TRANSFERABILITY ON MEDICAL ENDOSCOPIC DATASET IN TERMS OF MIOU, WHERE THE LAST ROW IS THE p VALUE OF T-TEST OF COMPARISON METHODS.

	ADV [39]	SIBAN [57]	DPR [58]	Ours
without quantified transferability	63.05	61.92	63.24	**64.17**
with quantified transferability	63.41	62.24	63.52	**64.43**
t-test (p value)	0.0014	0.0023	0.0042	0.0031

presented results in Fig. 4, we can notice that the performance degrades heavily after removing one or more components. Specifically, the mIoU drops 0.38%∼4.42% when pseudo labels selection, semantic representation transfer or quantified transferability is removed. It validates the designs of different components are reasonable. Compared with our conference version [3], the quantified transferability module and self-supervised pseudo labels generator are indispensable to boost the semantic knowledge transfer across domains. It further justifies that our model can effectively incorporate confident pseudo labels and transferable dependencies to improve segmentation performance.

2) Qualitative Analysis About Pseudo Labels: To verify the importance of our self-supervised pseudo label generator with class balance and spatial continuity, we conduct comparison experiments with other ablation designs, *i.e.*, optimizing without pseudo labels (Our-woPL), optimizing without super-pixel spatial prior (Ours-woSP), optimizing without class balance (Ours-woCB) and optimizing without self-supervision (Ours-woSS). As shown in Fig. 5, compared with Ours-woPL, our proposed pseudo label generator achieves 2.36% improvement in terms of mIoU by incorporating class balance and spatial prior in a self-supervised manner. Besides, it performs better than our conference version SWE [3], since the quantified transferability module and self-supervised training manner can explore transferable representations and prevent deviation of false pseudo labels in the retraining phase. Furthermore, Fig. 6 intuitively demonstrates the propagation process of pseudo labels, where our model could iteratively produce more confident pseudo labels for fine-tuning process.

3) Effect of Hyper-Parameters: We intend to study the parameter effects with respect to $\{\mu, \eta\}$ and $\{\alpha, \gamma\}$ in this subsection. The optimal values of hyper-parameters are determined by empirically conducting extensive parameter experiments, as presented in Fig. 7. We set $\eta = 0.3, \mu = 10, \alpha = 1$ and $\gamma = 0.7$ for best performance in all experiments. More-

TABLE IV
COMPARISON EXPERIMENTS ON SYNTHIA [56] → CITYSCAPES [54] TASK, WHERE mIoU* DENOTES THE mIoU FOR THE REMAINING 13 CLASSES.

Method	road	sidewalk	building	wall	fence	pole	light	sign	veg	sky	person	rider	car	bus	mbike	bike	mIoU	mIoU*
Wild [25]	11.5	19.6	30.8	4.4	0.0	20.3	0.1	11.7	42.3	68.7	51.2	3.8	54.0	3.2	0.2	0.6	20.2	-
CL [23]	65.2	26.1	74.9	0.1	0.5	10.7	3.7	3.0	76.1	70.6	47.1	8.2	43.2	20.7	0.7	13.1	29.0	-
NMD [26]	62.7	25.6	78.3	-	-	-	1.2	5.4	81.3	81.0	37.4	6.4	63.5	10.1	1.2	4.6	-	35.3
LSD [30]	80.1	29.1	77.5	2.8	0.4	26.8	11.1	18.0	78.1	76.7	48.2	15.2	70.5	17.4	8.7	16.7	36.1	-
LtA [33]	84.3	42.7	77.5	-	-	-	4.7	7.0	77.9	82.5	54.3	21.0	72.3	32.2	18.9	32.3	-	46.7
MCD [29]	84.8	**43.6**	79.0	3.9	0.2	29.1	7.2	5.5	83.8	83.1	51.0	11.7	79.9	27.2	6.2	0.0	37.3	-
CGAN [31]	85.0	25.8	73.5	3.4	**3.0**	31.5	19.5	21.3	67.4	69.4	**68.5**	25.0	76.5	**41.6**	17.9	29.5	41.2	-
DCAN [28]	82.8	36.4	75.7	5.1	0.1	25.8	8.0	18.7	74.7	76.9	51.1	15.9	77.7	24.8	4.1	37.3	38.4	-
CBST [34]	53.6	23.7	75.0	12.5	0.3	**36.4**	23.5	26.3	**84.8**	74.7	67.2	17.5	**84.5**	28.4	15.2	**55.8**	42.5	-
CLAN [40]	81.3	37.0	80.1	-	-	-	16.1	13.7	78.2	81.5	53.4	21.2	73.0	32.9	22.6	30.7	-	47.8
SWD [38]	82.4	33.2	**82.5**	-	-	-	22.6	19.7	83.7	78.8	44.0	17.9	75.4	30.2	14.4	39.9	-	48.1
ADV [39]	**85.6**	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4	**84.1**	57.9	23.8	73.3	36.4	14.2	33.0	41.2	-
DPR [58]	82.4	38.0	78.6	8.7	0.6	26.0	3.9	11.1	75.5	84.6	53.5	21.6	71.4	32.6	19.3	31.7	40.0	-
SIBAN [57]	82.5	24.0	79.4	-	-	-	16.5	12.7	79.2	82.8	58.3	18.0	79.3	25.3	17.6	25.9	-	46.3
MSL [59]	82.9	40.7	80.3	10.2	0.8	25.8	12.8	18.2	82.5	82.2	53.1	18.0	79.0	31.4	10.4	35.6	41.4	-
CRST-LR [35]	65.6	30.3	74.6	13.8	1.5	35.8	23.1	29.1	77.0	77.5	60.1	28.5	82.2	22.6	20.1	41.9	42.7	-
CRST-MR [35]	63.4	27.1	76.4	14.2	1.4	35.2	23.6	**29.4**	78.5	77.8	61.4	**29.5**	82.2	22.8	18.9	42.3	42.8	-
SWE [3]	68.4	30.1	74.2	21.5	0.4	29.2	29.3	25.1	80.3	81.5	63.1	16.4	75.6	13.5	26.1	51.9	42.9	-
BL	22.5	15.4	74.1	9.2	0.1	24.6	6.6	11.7	75.0	82.0	56.5	18.7	34.0	19.7	17.1	18.5	30.4	34.8
BL+AL	74.4	30.5	75.8	13.2	0.2	19.7	4.4	4.9	78.2	82.7	44.4	16.0	63.2	33.3	13.5	26.2	36.3	42.1
BL+AL+PL	79.5	39.4	77.3	11.2	0.3	22.5	5.3	11.8	79.2	81.8	58.6	20.3	71.2	31.0	25.3	33.1	40.5	47.2
BL+AL+SRT	79.9	38.2	77.1	9.7	0.2	21.1	6.8	7.6	76.1	81.6	54.8	21.3	66.2	30.8	21.6	30.6	39.0	45.6
BL+PL+SRT	62.8	28.5	71.3	21.3	0.6	29.2	32.1	25.3	80.7	81.1	62.3	15.7	68.9	12.6	27.6	51.7	42.0	47.8
BL+AL+PL+QT	80.3	39.6	76.8	13.7	0.5	22.3	5.8	12.1	78.7	82.2	58.4	19.6	71.9	31.3	26.5	34.1	40.9	47.5
BL+AL+SRT+QT	80.7	38.5	77.4	9.2	0.4	20.7	6.5	7.3	78.5	82.3	54.1	21.8	67.5	31.7	21.2	32.4	39.4	46.2
BL+PL+SRT+QT	67.3	31.5	76.9	21.4	0.5	28.1	31.9	24.2	78.4	80.2	61.4	15.9	70.9	13.2	24.1	49.2	42.2	48.1
Ours-woSP	70.8	32.9	76.1	21.6	0.6	28.9	30.5	23.6	77.2	78.8	62.3	16.4	72.5	12.1	**28.6**	49.4	42.6	48.5
Ours	70.3	32.1	75.7	**22.9**	0.8	29.6	**32.4**	24.3	79.7	78.1	62.5	17.0	75.0	13.2	28.3	51.7	**43.4**	**49.3**

over, our proposed model achieves stable transfer performance on the medical endoscopic dataset even though there is a wide selection range for hyper-parameters. The results of Fig. 7 (b) justify that incorporating previously-learned features and sparsity property for different feature centroids is essential to improve the performance of our model.

4) Effect of Quantified Transferability: This subsection validates the effectiveness of quantified transferability when combined with other state-of-the-art models, such as ADV [39], SIBAN [57] and DPR [58], as presented in Table III. From the presented results, we observe that: 1) The performance of comparison methods equipped with our proposed quantified transferability improves 0.28%∼0.36% mIoU on medical endoscopic datasets. It illustrates that the quantified transferability module could explore transferable knowledge while preventing negative transfer of untransferable knowledge. 2) Our model outperforms these competing methods even with quantified transferability about 0.91% ∼ 2.19% mIoU, which further verifies the superiority of our model.

To further validate the effectiveness of quantified transferability, we conduct the t-test experiments on all datasets to evaluate whether the improvement with quantified transferability mechanism is significant, as shown in Table III. Specifically, the lower value of p in Table III is, the more confident the significant improvement between the comparison methods with quantified transferability will be. We run each comparison method for 5 times, and then conduct the t-test evaluation. Obviously, the presented results (*i.e.*, p<0.05) effectively demonstrate the significance of the improvements of comparison methods with quantified transferability.

5) Convergence Analysis: This subsection introduces the convergence analysis about our proposed model. As shown in Fig. 8, our model could achieve efficient convergence (*i.e.*, iteration number is about 3.5×10^4) along the iterative training process. The experimental convergence curve effectively supports the theoretical analysis in Section III-E, *i.e.*, our model could narrow the domain discrepancy $d_{\mathcal{H}\triangle\mathcal{H}(P_s,P_t)}$ and minimize the upper bound of target expected error $\epsilon_T(h)$.

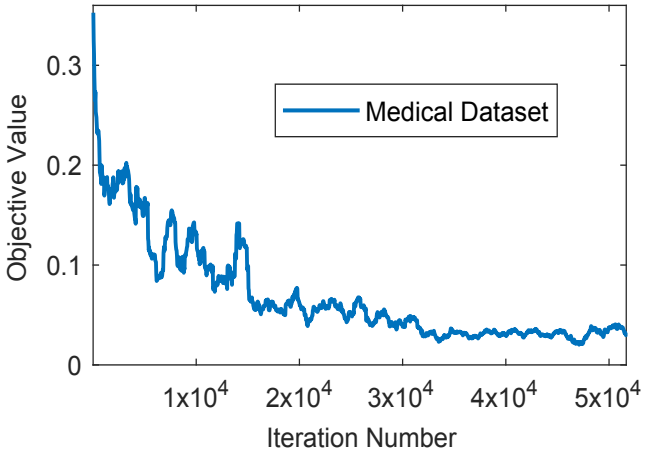


Fig. 8. Convergence analysis on medical endoscopic dataset.

C. Experiments on Benchmark Datasets

In this subsection, to further verify the generalization and robustness of our model, we conduct extensive comparison experiments on several representative public datasets (*i.e.*, SYNTHIA [56] → Cityscapes [54] and GTA [55] → Cityscapes [54]). To be specific, we remove the image classifier to neglect its assistance for pixel predictions, and utilize the same experiment settings with other comparison approaches [3], [30], [31], [33], [38], [39].

TABLE V
COMPARISON EXPERIMENTS ON GTA [55] → CITYSCAPES [54] TRANSFER TASK.

Method	road	sidewalk	building	wall	fence	pole	light	sign	veg	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
Wild [25]	70.4	32.4	62.1	14.9	5.4	10.9	14.2	2.7	79.2	21.3	64.6	44.1	4.2	70.4	8.0	7.3	0.0	3.5	0.0	27.1
CL [23]	74.9	22.0	71.7	6.0	11.9	8.4	16.3	11.1	75.7	11.3	66.5	38.0	9.3	55.2	18.8	18.9	0.0	16.8	14.6	28.9
CyCADA [27]	79.1	33.1	77.9	23.4	17.3	32.1	33.3	31.8	81.5	26.7	69.0	62.8	14.7	74.5	20.9	25.6	6.9	18.8	20.4	39.5
LSD [30]	88.0	30.5	78.6	25.2	23.5	16.7	23.5	11.6	78.7	27.2	71.9	51.3	19.5	80.4	19.8	18.3	0.9	20.8	18.4	37.1
LtA [33]	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
MCD [29]	90.3	31.0	78.5	19.7	17.3	28.6	30.9	16.1	83.7	30.0	69.1	58.5	19.6	81.5	23.8	30.0	5.7	25.7	14.3	39.7
CGAN [31]	89.2	49.0	70.7	13.5	10.9	38.5	29.4	33.7	77.9	**37.6**	65.8	**75.1**	**32.4**	77.8	**39.2**	45.2	0.0	25.2	35.4	44.5
DLOW [37]	87.1	33.5	80.5	24.5	13.2	29.8	29.5	26.6	82.6	26.7	**81.8**	55.9	25.3	78.0	33.5	38.7	0.0	22.9	34.5	42.3
CBST [34]	88.0	**56.2**	77.0	27.4	22.4	40.7	47.3	40.9	82.4	21.6	60.3	50.2	20.4	83.8	35.0	**51.0**	15.2	20.6	37.0	46.2
CLAN [40]	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2
SWD [38]	92.0	46.4	**82.4**	24.8	24.0	35.1	33.4	34.2	83.6	30.4	80.9	56.9	21.9	82.0	24.4	28.7	6.1	25.0	33.6	44.5
ADV [39]	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	**83.9**	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.5	45.5
DPR [58]	92.3	51.9	82.1	29.2	25.1	24.5	33.8	33.0	82.4	32.8	82.2	58.6	27.2	84.3	33.4	46.3	2.2	29.5	32.3	46.5
SIBAN [57]	88.5	35.4	79.5	26.3	24.3	28.5	32.5	18.3	81.2	40.0	76.5	58.1	25.8	82.6	30.3	34.4	3.4	21.6	21.5	42.6
MSL [59]	89.4	43.0	82.1	30.5	21.3	30.3	34.7	24.0	85.3	39.4	78.2	63.0	22.9	84.6	36.4	43.0	5.5	34.7	33.5	46.4
CRST-LR [35]	80.3	40.8	65.8	24.6	30.5	43.1	49.5	40.3	82.1	26.0	54.6	59.4	32.1	68.0	31.9	30.0	21.9	**44.8**	**46.7**	45.9
CRST-MR [35]	84.4	52.7	74.7	**38.0**	**32.2**	**43.7**	**53.7**	38.6	73.9	24.4	64.4	45.6	24.6	63.2	3.22	31.9	**45.9**	44.2	34.8	46.0
SWE [3]	92.7	48.0	78.8	25.7	27.2	36.0	42.2	45.3	80.6	14.6	66.0	62.1	30.4	86.2	28.0	45.6	35.9	16.8	34.7	47.2
BL	80.2	6.4	74.8	8.8	17.2	17.5	30.5	17.7	75.0	14.1	57.9	56.2	27.3	64.1	29.7	24.1	4.7	27.6	33.4	35.1
BL+AL	86.3	32.2	79.8	22.0	22.2	27.1	33.5	20.1	80.3	21.5	75.5	59.0	25.4	73.1	28.0	32.2	5.4	27.3	31.5	41.2
BL+AL+PL	91.3	48.7	77.6	23.9	28.7	30.4	41.8	43.0	77.8	15.4	73.5	55.6	19.8	86.3	22.5	43.2	34.5	14.6	30.8	45.2
BL+AL+SRT	92.4	49.8	73.6	25.3	28.3	24.5	40.9	45.0	79.2	14.2	70.4	50.1	18.6	86.6	22.3	45.4	30.3	11.9	32.8	44.3
BL+PL+SRT	92.4	48.3	75.6	27.5	27.3	30.5	42.7	45.9	81.7	15.8	73.0	56.5	25.8	86.1	23.6	41.7	43.8	13.3	30.9	46.4
BL+AL+PL+QT	92.1	49.4	78.3	22.6	26.1	32.7	42.3	41.8	79.5	16.1	72.7	57.8	17.4	85.6	24.8	42.5	36.2	15.1	31.4	45.5
BL+AL+SRT+QT	92.2	50.4	75.1	24.6	27.5	27.2	39.6	44.1	80.3	15.7	72.8	55.2	16.5	85.3	21.4	43.9	33.6	12.3	31.1	44.7
BL+PL+SRT+QT	92.5	48.3	73.8	27.8	26.1	28.9	44.3	45.2	80.4	17.6	73.5	62.7	24.1	84.7	27.0	42.6	41.5	16.4	31.3	46.8
Ours-woSP	92.5	47.9	77.1	24.9	27.3	35.4	41.5	43.8	78.2	18.7	72.4	64.3	31.5	83.8	27.9	43.4	32.7	18.5	33.6	47.1
Ours	**92.8**	48.6	79.2	26.2	26.4	36.8	40.6	**47.1**	80.5	19.3	69.0	63.1	28.9	**86.7**	29.2	45.8	34.3	19.4	32.7	**47.7**

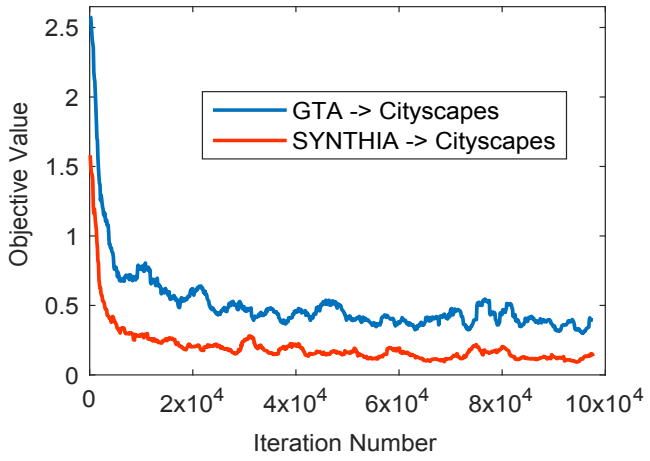


Fig. 9. Convergence analysis of our model on GTA → Cityscapes and SYNTHIA → Cityscapes adaptation tasks.

TABLE VI
THE EFFECT OF QUANTIFIED TRANSFERABILITY ON SYNTHIA → CITYSCAPES (MIOU), WHERE THE LAST ROW IS THE p VALUE OF T-TEST.

	ADV [39]	SIBAN [57]	DPR [58]	Ours
without quantified transferability	41.2	37.4	40.0	**43.1**
with quantified transferability	41.6	37.7	40.5	**43.4**
t-test (p value)	0.0051	0.0063	0.0042	0.0048

The convergence analysis of our model on benchmark datasets are depicted in Fig. 9, which validates that our model achieves tight upper bound for target expected error $\epsilon_T(h)$ along the iterative training process. The experimental results in Fig. 9 well match the theoretical analysis in Section III-E.

1) Transfer from SYNTHIA to Cityscapes: Learning transferable dependencies from SYNTHIA [56] to Cityscapes [54] is our focus in this experiment. In the training stage, we regard finely-labeled 9400 samples from SYNTHIA and 2993 samples from Cityscapes without pixel labels as X^s and X^t, respectively. For the evaluation, validation subset with 500 images from Cityscapes dataset is utilized, where this subset is disjoint with the training data. As shown in Table IV, 16 shared classes for SYNTHIA and Cityscapes datasets are considered. From Table IV, we can notice that: 1) Our model improves the transfer performance of $0.4\% \sim 6.0\%$ than competing models [29], [34], [38]–[40] and our conference version [3] for 16 categories in terms of mIoU, which validates our proposed model can efficiently explore transferable dependencies across source and target datasets while neglecting the irrelevant knowledge; 2) Ablation performance further justifies each designed component play an important role in narrowing the semantic gap; 3) When compared with [3], our model has comparable performance for hard-to-transfer categories with extremely different appearances across domains (*e.g.*, wall, pole, light, motorbike and rider), since pseudo label generator generates confident pixel labels for target samples by considering both class balance and super-pixel prior in a self-supervised manner. Furthermore, as presented in Table VI, the quantified transferability module is combined with some state-of-the-art models, *i.e.*, ADV [39], SIBAN [57] and DPR

TABLE VII
THE EFFECT OF QUANTIFIED TRANSFERABILITY ON GTA → CITYSCAPES IN TERMS OF MIOU, WHERE THE LAST ROW IS THE p VALUE OF T-TEST.

	ADV [39]	SIBAN [57]	DPR [58]	Ours
without quantified transferability	45.5	42.6	46.5	**47.5**
with quantified transferability	45.9	42.8	46.8	**47.7**
t-test (p value)	0.0047	0.0084	0.0061	0.0096

[58] to further demonstrate its scale application ability for other models. From the t-test evaluation results in Table VI, we could observe that the competing models with quantified transferability improve the transfer performance significantly.

2) Transfer from GTA to Cityscapes: For model training, we employ GTA [55] with finely-annotated 24996 samples as source dataset X^s, and utilize the training subset of Cityscapes [54] with unlabeled 2993 samples as target dataset X^t. Besides, we use the validation subset of Cityscapes [54] including 500 samples to evaluate the performance. Table V reports the results for 19 common classes. From Table V, we have the following observations: 1) Our model pays equal attention for both easy-to-transfer categories and initially hard-to-transfer classes, while other competing methods [25], [30], [31], [33] give priority to those well-to-transfer classes with large amount of pixels (*e.g.*, vegetation, building, road, car and sky). 2) Ablation experiments validate that each developed module reasonable designs are efficient to improve semantic transfer performance. 3) Compared with our conference paper [3], our model with the Wasserstein quantified transferability focuses on those transferable semantic knowledge and simultaneously ignores the irrelevant regions (*e.g.*, background and some invalid classes). 4) The self-supervised pseudo label generator produces highly-confident pseudo labels, which assists the feature distributions among source and target datasets in both global and local views aligned better. Moreover, as shown in Table VII, we apply the quantified transferability module into other competing models, *i.e.*, ADV [39], SIBAN [57] and DPR [58] to further validate its application advantage and the significant improvement for knowledge transfer performance.

V. Conclusion

In this paper, we present a new weakly-supervised lesions transfer framework, which highlights important transferable dependencies while neglecting those irrelevant untransferable representations. To be specific, a Wasserstein adversarial learning based quantified transferability framework is developed to selectively capture transferable semantic representations across different domains. Moreover, we propose a self-supervised pseudo pixel label generator under curriculum learning strategy to pick more hard-to-transfer samples from target domain for the model retraining process. It incorporates class balance and super-pixel prior to prevent the enormous deviation of false pseudo labels and the dominance of well-to-transfer classes in a self-supervised manner. With the assistance of dynamically-updated semantic consistence loss, category-wise feature centroids are aligned to narrow the conditional distributions across domains. Extensive comparison experiments on several medical and non-medical domain adaptation datasets justify the robustness and effectiveness of our model. In the future, we will explore unsupervised semantic transfer across gastroscope and enteroscopy datasets for lesions segmentation and apply it into clinical diagnosis applications.

References

[1] Z. Jia, X. Huang, E. I.-C. Chang, and Y. Xu, "Constrained Deep Weak Supervision for Histopathology Image Segmentation," *IEEE Transactions on Medical Imaging*, vol. 36, pp. 2376–2388, Nov. 2017.

[2] Y. Xu, J.-Y. Zhu, E. I.-C. Chang, M. Lai, and Z. Tu, "Weakly supervised histopathology cancer image segmentation and classification," *Medical Image Analysis*, vol. 18, pp. 591–604, Apr. 2014.

[3] J. Dong, Y. Cong, G. Sun, and D. Hou, "Semantic-transferable weakly-supervised endoscopic lesions segmentation," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[4] C. Yan, J. Yao, R. Li, Z. Xu, and J. Huang, "Weakly supervised deep learning for thoracic disease classification and localization on chest x-rays," in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2018, pp. 103–110.

[5] R. Zhao, W. Liao, B. Zou, Z. Chen, and S. Li, "Weakly-supervised simultaneous evidence identification and segmentation for automated glaucoma diagnosis," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 809–816, Jul. 2019.

[6] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[7] K. Doi, "Computer-aided diagnosis in medical imaging: Historical review, current status and future potential," *Computerized Medical Imaging and Graphics*, vol. 31, no. 4-5, pp. 198–211, Jun. 2007.

[8] K. Horsch, M. L. Giger, L. A. Venta, and C. J. Vyborny, "Automatic segmentation of breast lesions on ultrasound," *Medical Physics*, vol. 28, no. 8, pp. 1652–1659, Aug. 2001.

[9] S.-C. Cheng, K.-Y. Cheng, and Y.-P. P. Chen, "Ght-based associative memory learning and its application to human action detection and classification," *Pattern Recognition*, vol. 46, no. 11, pp. 3117–3128, 2013.

[10] B. Lin, Y. Sun, J. E. Sanchez, and X. Qian, "Efficient vessel feature detection for endoscopic image analysis," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 4, pp. 1141–1150, 2015.

[11] J.-Z. Cheng, Y.-H. Chou, C.-S. Huang, Y.-C. Chang, C.-M. Tiu, K.-W. Chen, and C.-M. Chen, "Computer-aided US Diagnosis of Breast Lesions by Using Cell-based Contour Grouping," *Radiology*, vol. 255, no. 3, pp. 746–754, Jun. 2010.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[13] Q. Wang, H. Fan, G. Sun, Y. Cong, and Y. Tang, "Laplacian pyramid adversarial network for face completion," *Pattern Recognition*, vol. 88, pp. 493–505, 2019.

[14] Q. Wang, H. Fan, G. Sun, W. Ren, and Y. Tang, "Recurrent generative adversarial network for face completion," *IEEE Transactions on Multimedia*, 2020.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing, 2015, pp. 234–241.

[16] Z. Jiang, H. Zhang, Y. Wang, and S.-B. Ko, "Retinal blood vessel segmentation using fully convolutional network with transfer learning," *Computerized Medical Imaging and Graphics*, vol. 68, pp. 1–15, Sep. 2018.

[17] D. Pathak, P. Krahenbuhl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV '15. USA: IEEE Computer Society, 2015, pp. 1796–1804.

[18] H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov, and I. B. Ayed], "Constrained-cnn losses for weakly supervised segmentation," *Medical Image Analysis*, vol. 54, pp. 88 – 99, 2019.

[19] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, 2015, pp. 1180–1189.

[20] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *arXiv preprint arXiv:1511.06434*, Nov. 2016.

[21] V. K. Kurmi, S. Kumar, and V. P. Namboodiri, "Attending to discriminative certainty for domain adaptation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[22] J. Li, M. Jing, K. Lu, L. Zhu, and H. T. Shen, "Locality preserving joint transfer for domain adaptation," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 6103–6115, Dec 2019.

[23] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[24] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 95–104.

[25] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "Fcns in the wild: Pixel-level adversarial and constraint-based adaptation," *arXiv preprint arXiv:1612.02649*, 2016.

[26] Y.-H. Chen, W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y.-C. Frank Wang, and M. Sun, "No more discrimination: Cross city adaptation of road scene segmenters," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[27] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, Jul 2018, pp. 1989–1998.

[28] Z. Wu, X. Han, Y.-L. Lin, M. Gokhan Uzunbas, T. Goldstein, S. Nam Lim, and L. S. Davis, "Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[29] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[30] S. Sankaranarayanan, Y. Balaji, A. Jain, S. Nam Lim, and R. Chellappa, "Learning from synthetic data: Addressing domain shift for semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[31] W. Hong, Z. Wang, M. Yang, and J. Yuan, "Conditional generative adversarial network for structured domain adaptation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[32] H. Xia and Z. Ding, "Structure preserving generative cross-domain learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[33] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[34] Y. Zou, Z. Yu, B. Vijaya Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[35] Y. Zou, Z. Yu, X. Liu, B. V. Kumar, and J. Wang, "Confidence regularized self-training," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[36] Q. Wang, J. Gao, and X. Li, "Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes," *IEEE Transactions on Image Processing*, vol. 28, pp. 4376–4386, 2019.

[37] R. Gong, W. Li, Y. Chen, and L. V. Gool, "Dlow: Domain flow for adaptation and generalization," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[38] C.-Y. Lee, T. Batra, M. H. Baig, and D. Ulbricht, "Sliced wasserstein discrepancy for unsupervised domain adaptation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[39] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Perez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[40] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[41] J. Dong, Y. Cong, G. Sun, Y. Liu, and X. Xu, "Cscl: Critical semantic-consistent learning for unsupervised domain adaptation," in *European Conference on Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 745–762.

[42] J. Dong, Y. Cong, G. Sun, B. Zhong, and X. Xu, "What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 4022–4031.

[43] W. Deng, L. Zheng, Y. Sun, and J. Jiao, "Rethinking triplet loss for domain adaptation," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 01 2020.

[44] X. Xu, H. He, H. Zhang, Y. Xu, and S. He, "Unsupervised domain adaptation via importance sampling," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2019.

[45] L. Zhang, P. Wang, W. Wei, H. Lu, C. Shen, A. van den Hengel, and Y. Zhang, "Unsupervised domain adaptation using robust class-wise matching," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 5, pp. 1339–1349, 2019.

[46] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 214–223.

[47] H. Zhang, I. J. Goodfellow, D. N. Metaxas, and A. Odena, "Self-attention generative adversarial networks," *arXiv preprint arXiv:1805.08318*, 2018.

[48] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[49] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[50] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 834–848, Apr. 2018.

[51] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[52] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, pp. 151–175, May 2010.

[53] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 1857–1866.

[54] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[55] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *The IEEE International Conference on Computer Vision (ICCV)*, 2016.

[56] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[57] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Significance-aware information bottleneck for domain adaptive semantic segmentation," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[58] Y.-H. Tsai, K. Sohn, S. Schulter, and M. Chandraker, "Domain adaptation for structured output via discriminative patch representations," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[59] M. Chen, H. Xue, and D. Cai, "Domain adaptation for semantic segmentation with maximum squares loss," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.



Jiahua Dong is currently a Ph. D candidate in State Key Laboratory of Robotics, Shenyang Institute of Automation, University of Chinese Academy of Sciences. He received the B.S. degree from Jilin University in 2017. His current research interests include computer vision, machine learning, transfer learning, domain adaptation and medical image processing.



Yang Cong (S'09-M'11-SM'15) is a full professor of Chinese Academy of Sciences. He received the he B.Sc. de. degree from Northeast University in 2004, and the Ph.D. degree from State Key Laboratory of Robotics, Chinese Academy of Sciences in 2009. He was a Research Fellow of National University of Singapore (NUS) and Nanyang Technological University (NTU) from 2009 to 2011, respectively; and a visiting scholar of University of Rochester. He has served on the editorial board of the Journal of Multimedia. His current research interests include image processing, compute vision, machine learning, multimedia, medical imaging, data mining and robot navigation. He has authored over 70 technical papers. He is also a senior member of IEEE.



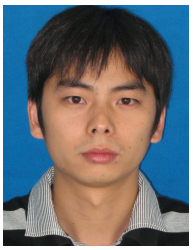
Gan Sun (S'19) is an Assistant Professor in State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences. He received the B.S. degree from Shandong Agricultural University in 2013, the Ph.D. degree from State Key Laboratory of Robotics, Chinese Academy of Sciences in 2020, and has been visiting Northeastern University from April 2018 to May 2019, Massachusetts Institute of Technology from June 2019 to November 2019. His current research interests include lifelong machine learning, multi-task learning, medical data analysis, deep learning and 3D computer vision.



Yunsheng Yang is currently a professor of Department of Gastroenterology and Hepatology, director of his department for 12 years and Chairman for the Institute of Digestive Diseases in the Chinese PLA General Hospital (301 Hospital) in Beijing. He serves as the supervisor for Ph.D student at Nankai University in Tianjing, Tsinghua University and Chinese PLA Medical Academy in Beijing. Professor Yang graduated from Medical College in 1983, received his master's degree of medical science from Henan Medical University in 1992 and Ph.D degree from the First Military Medical University in 1995. Professor Yang is the Founder of Sino-NOTES Club and also serves of various of editorial boards and editor-in-chief for Chinese Version of Gut and Chinese Version of Gastroenterology, associated editor-in-chief JDD and Chin J Digestion, editorial member of Gut, and APT, etc. So far, his more than 200 articles have been published, including 150 of original articles in various of international journals, like gastrointestinal endoscopy GIE, hepatology and so on.



Xiaowei Xu is a professor of Information Science at University of Arkansas at Little Rock (UALR), received a a B.Sc. de. degree in Mathematics from Nankai University in 1983 and a Ph.D. degree in Computer Science from University of Munich in 1998. He holds an adjunct professor position in the Department of Mathematics and Statistics at University of Arkansas at Fayetteville. Before his appointment in UALR, he was a senior research scientist in Siemens. He was a visiting professor in Microsoft Research Asia and Chinese University of Hong Kong. His research spans data mining, machine learning, bioinformatics, data management and high performance computing. He has published over 70 papers in peer reviewed journals and conference proceedings. His groundbreaking work on density-based clustering algorithm DBSCAN has been widely used in many textbooks; and received over 10203 citations based on Google scholar. Dr. Xu is a recipient of 2014 ACM KDD Test of Time Award that "recognizes outstanding papers from past KDD Conferences beyond the last decade that have had an important impact on the data mining research community."



Zhengming Ding (S'14-M'18) received the B.Eng. degree in information security and the M.Eng. degree in computer software and theory from University of Electronic Science and Technology of China (UESTC), China, in 2010 and 2013, respectively. He received the Ph.D. degree from the Department of Electrical and Computer Engineering, Northeastern University, USA in 2018. He is a faculty member affiliated with Department of Computer, Informationand Technology, Indiana University-Purdue University Indianapolis since 2018. His research interests include transfer learning, multi-view learning and deep learning. He received the National Institute of Justice Fellowship during 2016-2018. He was the recipients of the best paper award (SPIE 2016) and best paper candidate (ACMMM 2017). He is currently an Associate Editor of the Journal of Electronic Imaging (JEI) and IET Image Processing. He is a member of IEEE, ACM and AAAI.