Where and How to Transfer: Knowledge Aggregation-Induced Transferability Perception for Unsupervised Domain Adaptation

Jiahua Dong, Yang Cong, *Senior Member, IEEE,* Gan Sun, Zhen Fang and Zhengming Ding

Abstract—Unsupervised domain adaptation without accessing expensive annotation processes of target data has achieved remarkable successes in semantic segmentation. However, most existing state-of-the-art methods cannot explore whether semantic representations across domains are transferable or not, which may result in the negative transfer brought by irrelevant knowledge. To tackle this challenge, in this paper, we develop a novel Knowledge Aggregation-induced Transferability Perception (KATP) module for unsupervised domain adaptation, which is a pioneering attempt to distinguish transferable or untransferable knowledge across domains. Specifically, the KATP module is designed to quantify which semantic knowledge across domains is transferable, by incorporating the transferability information propagation from constructed global category-wise prototypes. Based on KATP, we design a novel KATP Adaptation Network (KATPAN) to determine where and how to transfer. The KATPAN contains a transferable appearance translation module $\mathcal{T}_A(\cdot)$ and a transferable representation augmentation module $\mathcal{T}_R(\cdot)$, where both modules construct a virtuous circle of performance promotion. $\mathcal{T}_A(\cdot)$ develops a transferability-aware information bottleneck to highlight where to adapt transferable visual characterizations and modality information; $\mathcal{T}_R(\cdot)$ explores how to augment transferable representations while abandoning untransferable information, and promotes the translation performance of $\mathcal{T}_A(\cdot)$ in return. Comprehensive experiments on several representative benchmark datasets and a medical dataset support the state-of-the-art performance of our model.

Index Terms—Transfer Learning, Unsupervised Domain Adaptation, Semantic Segmentation, Medical Lesions Diagnosis.

1 Introduction

Convolutional neural networks (CNNs) have shown great ability to characterize semantic context and are successfully applied into semantic segmentation tasks [3], [18], [36], [38], [50]. Unfortunately, the success of CNNs consumes the large quantity of labor efforts to manually collect and annotate the training data with dense pixel-level annotations. Furthermore, the performance of learned model undergoes a significant decrease, when the training and testing datasets are drawn from different domains (*i.e.*, domain gap). Consequently, unsupervised domain adaptation methods [24]–[27], [42] have been used to narrow the domain gap in the semantic segmentation tasks. Specifically, existing methods mainly employ adversarial framework [5], [13], [22], [27], [43] in the feature space or use pseudo label generation strategy [10], [11], [26], [54] for target samples to extract domain-invariant knowledge from annotated source domain and unlabeled target domain. Using such domain-invariant knowledge, they can learn effective pixel-level classifiers that have good prediction performance on the unlabeled target domain [9], [10], [32], [49].

However, existing methods assume that semantic representations have the same degree of contributions for the domain adaptation procedure [13], [27], [41], [43], [47], which is not realistic. Taking the urban scene segmentation as an example, the objects (*e.g.*, road and sky) with a large number of pixels are the easy-to-adapt classes across domains, which have different significances of knowledge transfer with the uncommon hard-to-adapt categories (*e.g*, pole and light). It is a challenging task to manually quantify the contributions of each object across domains. Furthermore, the categories (*e.g.*, bus, car, person, etc) with diverse appearances and different modality information across source and target domains significantly mislead most existing advanced methods [10], [13], [27] to undergo the negative transfer of irrelevant knowledge. Consequently, automatically highlighting transferable visual characterizations and semantic representations while preventing the negative transfer brought by untransferable knowledge is a crucial challenge for domain adaptation task.

In this paper, we investigate that semantic representations have different degrees of contributions for the domain

- *Jiahua Dong is with the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, 110016, China, also with the Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang, 110169, China, and also with the University of Chinese Academy of Sciences, Beijing, 100049, China. Email: dongjiahua1995@gmail.com.*
- *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, 110169, China. Email: congyang81@gmail.com, sungan1412@gmail.com.*
- *Zhen Fang is with the Australian Artificial Intelligence Institute, University of Technology Sydney, NSW 2007, Australia. Email: Zhen.Fang@student.uts.edu.au.*
- *Zhengming Ding is with the Department of Computer Science, Tulane University, New Orleans, LA 70118, USA. Email: zding1@tulane.edu.*

Manuscript received April 19, 2005; revised August 26, 2015.
This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFB1310300; in part by the National Nature Science Foundation of China under Grant 61821005, and Grant 62003336; and in part by National Postdoctoral Innovative Talents Support Program (BX20200353).
The corresponding author is Prof. Yang Cong.

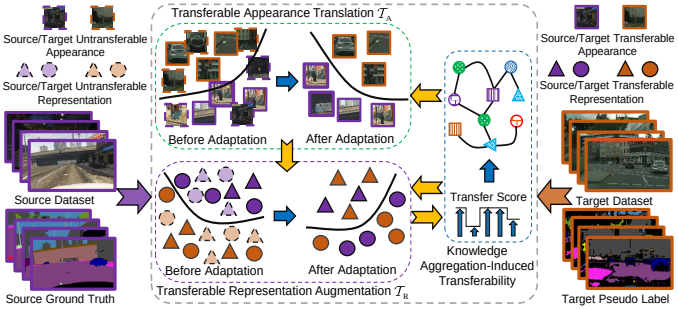


Fig. 1. Demonstration of our proposed KATPAN, where the knowledge aggregation-induced transferability perception (KATP) module quantifies which semantic knowledge across domains is transferable, two complementary modules $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$ respectively highlight where to purify transferable visual appearances and how to augment transferable semantic representations while preventing irrelevant knowledge transfer.

adaptation procedure in the real world (see the above examples). More importantly, some semantic features might even cause the negative transfer due to the irrelevant knowledge (see Table 8). To address the above issues, as shown in Fig. 1, we design a novel Knowledge Aggregation-induced Transferability Perception (KATP) module to identify which semantic representations across domains could be effectively transferred. Based on KATP, we propose a novel KATP Adaptation Network (KATPAN) that consists of a transferable appearance translation module $\mathcal{T}_A(\cdot)$ and a transferable representation augmentation module $\mathcal{T}_R(\cdot)$. Both modules respectively explore where to highlight transferable visual characterizations and how to augment transferable semantic representations across domains, while brushing untransferable and irrelevant knowledge aside.

To be specific, the KATP module focuses on quantifying the contributions of semantic representations for knowledge adaptation by exploring underlying transferability information propagation from global category-wise prototypes. We construct a transferability knowledge graph to effectively propagate the transferability of these prototypes to their adjacent areas. The $\mathcal{T}_A(\cdot)$ concentrates on identifying where to translate transferable visual appearances and modality information while abandoning the untransferable translation of visual characterizations, by incorporating with the transferability-aware information bottleneck. The $\mathcal{T}_R(\cdot)$ introduces a conditional feature augmentor with multiple residual attention on attention blocks and a self-adaptive pseudo label selection strategy to determine how to augment transferable representations across domains, while neglecting the irrelevant knowledge. Moreover, $\mathcal{T}_R(\cdot)$ improves the visual translation performance of $\mathcal{T}_A(\cdot)$ in return.

The superior performance of our proposed model is verified by extensive experiments on four representative benchmark datasets and a medical lesions dataset. The novel contributions of this paper are summarized as follows:

- We develop a novel Knowledge Aggregation-induced Transferability Perception Adaptation Network (KATPAN) for unsupervised domain adaptation. This is a pioneer exploration to distinguish transferable (or untransferable) semantic features and highlight (or neglect) corresponding knowledge in domain adaptation.
- A Knowledge Aggregation-induced Transferability Perception (KATP) module is designed to quantify which semantic knowledge across domains could be effectively transferred via the transferability information propagation from global category-wise prototypes.
- We propose two complementary modules $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$ to identify where to translate transferable visual characterizations and how to augment transferable representations across domains by establishing a virtuous circle of performance promotion.

This paper is a substantial extension of our conference version [12]. Compared with [12], several significant improvements of this paper could be summarized as follows: 1) We develop the KATP module to quantify which semantic representations across domains are transferable or not via the transferability information propagation from global category-wise prototypes. 2) We ameliorate the optimization manner of transferable representation augmentation module $\mathcal{T}_R(\cdot)$ to an end-to-end training manner, and significantly improve the performance of $\mathcal{T}_R(\cdot)$ by incorporating a conditional feature augmentor and a category-level adversarial framework with the quantified transferability perception. 3) A self-adaptive pseudo label selection strategy in $\mathcal{T}_R(\cdot)$ is designed to mine confident pseudo labels for target data. 4) Abundant comparison experiments on several representative datasets are conducted to verify the superiority of our proposed model against other competing methods.

2 Related Work

2.1 Semantic Segmentation

Recently, the remarkable successes for semantic segmentation [38] have mostly been achieved by deep convolutional network, due to the characterization ability for semantic representation. Ghiasi *et al.* introduce a Laplacian pyramid architecture [18] to integrate multi-scale semantic context. Different from [18], a fully convolutional neural network [38] is developed by Shelhamer *et al.* to perform dense pixel-level segmentation. [3] employs the atrous spatial pyramid pooling operation (*i.e.*, dilated convolution) to enlarge the receptive field. Some encoder-decoder networks (*e.g.*, PSANet [50], U-Net [36], etc) are proposed to fuse representations from different scales of semantic information. Liu *et al.* [30] employ Markov random field to explore high-order semantic relations and label context. [29] constructs pairwise relationship of semantic context via an affinity matrix. Unfortunately, they consume large-scale dense pixel-level annotations to achieve superior performance, which is expensive to manually annotate the collected data.

2.2 Unsupervised Domain Adaptation

After adversarial learning [53] is first introduced by Hoffman *et al.* to perform unsupervised domain adaptation task [10], [14], [15], [52], large quantities of competing variants [13], [27], [41]–[43], [46] relying on adversarial framework have been developed to explore shared domain-invariant features. Different from them, Zou *et al.* [54] employ the non-adversarial method to narrow the distribution discrepancy by mining confident class-balanced pseudo labels for target domain in a self-training manner. Furthermore, curriculum

learning is employed by [27], [48] to predict the category-wise distribution properties of target domain, which gradually bridges the domain-wise distribution discrepancy from the easy-to-adapt samples to the difficult ones. [26] proposes a bidirectional learning including the visual appearance translation and semantic feature adaptation to bridge the distribution shift across domains [28]. Motivated by [26], Dong *et al.* design two complementary modules with quantified transferability [12] to explore transferable appearances and features across domains. Moreover, Finn *et al.* [16] propose a model-agnostic meta-learning framework to perform fast adaptation for a variety of different applications that are optimized with gradient descent. Yu *et al.* [47] develop a fast online adaptive learning model to achieve a fast and robust adaptation for cardiac motion estimation by designing an online optimizer based on meta-learning. Some efficient objectives for knowledge adaptation [6] (*e.g.*, sliced Wasserstein discrepancy [24], maximum squares loss [4], adversarial entropy minimization [43] and scale-invariant formulation [41]) are designed to narrow the domain gap. In addition, Wang *et al.* [44] propose a fine-grained domain discriminator to distinguish different domains at category level. [32], [45] aim to transfer interactive relations among different objects, while our proposed model focuses on the transferability information propagation from global category-wise prototypes to quantify which semantic representations across domains could be transferred.

3 The Proposed KATPAN

Let $\mathcal{X}$ and $\mathcal{Y}$ be the input space and output space, respectively. In unsupervised domain adaptation, there are two different joint distributions defined over $\mathcal{X} \times \mathcal{Y}$: $\mathcal{P}_{\mathbf{X}_s\mathbf{Y}_s}$ (source domain) and $\mathcal{P}_{\mathbf{X}_t\mathbf{Y}_t}$ (target domain), where $\mathbf{X}_s, \mathbf{X}_t \in \mathcal{X}$ and $\mathbf{Y}_s, \mathbf{Y}_t \in \mathcal{Y}$ are random variables. Given labeled source data $X_s = \{x_i^s, y_i^s\}_{i=1}^{n_s} \sim \mathcal{P}_{\mathbf{X}_s\mathbf{Y}_s}$, i.i.d (independent and identically distributed), and unlabeled target data $X_t = \{x_j^t\}_{j=1}^{n_t} \sim \mathcal{P}_{\mathbf{X}_t}$, i.i.d, where n_s and n_t represent the number of samples in source and target datasets. In this paper, we focus on classifying the unlabeled target dataset without accessing the corresponding labels, in spite of the large domain discrepancy between source domain $\mathcal{P}_{\mathbf{X}_s\mathbf{Y}_s}$ and target domain $\mathcal{P}_{\mathbf{X}_t\mathbf{Y}_t}$. K denotes the number of classes.

3.1 Overview

As the graphical illustration depicted in Fig. 2, we develop a novel KATP Adaptation Network (KATPAN) to determine where and how to highlight transferable knowledge across domains while neglecting irrelevant knowledge. Specifically, with the KATP module to identify which semantic representations across domains could be effectively transferred, two complementary modules (*i.e.*, a transferable appearance translation module $\mathcal{T}_A(\cdot)$ and a transferable representation augmentation module $\mathcal{T}_R(\cdot)$) construct a virtuous circle of performance promotion to highlight where and how to transfer domain-invariant knowledge. Both source and target datasets are first forwarded into $\mathcal{T}_A(\cdot)$ to highlight where to capture transferable visual translation while preventing the negative transfer from untransferable appearances, and then fed into $\mathcal{T}_R(\cdot)$ to explore how to augment transferable knowledge while abandoning irrelevant representations.

3.2 KATP: A Knowledge Aggregation-induced Transferability Perception

3.2.1 Preliminary

In generative adversarial network based domain adaptation, as introduced in [12], the discriminator could assist in distinguishing the cross-domain semantic knowledge that is transferable or untransferable, by identifying whether a given sample is from source or target domain (*i.e.*, domain uncertainty estimation). For example, when the semantic representations across different domains are already adapted well, it could confuse the discriminator to distinguish which one is from source or target domain. Obviously, the output probability of discriminator plays an essential role in identifying whether the semantic representations across domains are transferable or untransferable. To this end, an uncertainty measure function from information theory (*i.e.*, entropy criterion $\mathcal{I}(p) = \sum_i p_i \log(p_i)$) is employed to quantify the transferability of corresponding representations across domains. Specifically, as shown in Fig. 2, given the feature $F_i^s \in \mathbb{R}^{H_s \times W_s \times U}$ of the i-th source sample, the discriminator $D_F(\cdot)$ with network parameters as θ_{D_F} takes F_i^s as input, and outputs the domain uncertainty probability $D_F(F_i^s; \theta_{D_F}) \in \mathbb{R}^{H_s \times W_s}$, where H_s, W_s and U respectively denote the height, width and channel dimensions of F_i^s. Then, the quantified transferability $Q_{F_s} \in \mathbb{R}^{H_s \times W_s}$ for source representation F_i^s can then be represented as follows:

$$Q_{F_s} = 1 - \mathcal{I}(D_F(F_i^s; \theta_{D_F})). \tag{1}$$

Similarly, given the domain uncertainty probability $D_F(F_j^t; \theta_{D_F}) \in \mathbb{R}^{H_t \times W_t}$, we also utilize the entropy criterion $\mathcal{I}(\cdot)$ to quantify the transferability $Q_{F_t} \in \mathbb{R}^{H_t \times W_t}$ for the representation $F_j^t \in \mathbb{R}^{H_t \times W_t \times U}$ of the j-th target image:

$$Q_{F_t} = 1 - \mathcal{I}(D_F(F_j^t; \theta_{D_F})), \tag{2}$$

where H_t, W_t and U represent the height, width and channel dimensions of F_j^t.

3.2.2 The Proposed KATP Module

However, in the real world, the transferability should be evaluated on each category rather than each pixel. For example, in the urban scene segmentation tasks, the categories with a large number of pixels (*e.g*, road and sky) are the easy-to-adapt categories, and all pixels regarding these categories should be assigned with higher transferability. If we only focus on evaluating transferability on each pixel (like Eqs. (1) and (2)), some pixels regarding these categories will be assigned lower transferability, which is not reasonable and thus causes false transferability perception.

To address the above issue, we develop a KATP module to evaluate the transferability of each pixel according to its category information. Specifically, we first construct global category-wise prototypes and quantify their transferability as reference. Then, the transferability of these prototypes is propagated to their adjacent areas via constructing a trasnferability knowledge graph. The category-wise prototypes ensure the accurate transferability of each category, and the proposed trasnferability knowledge graph ensures that semantically adjacent pixels have similar transferability.

• **Global Category-wise Prototypes Construction:** In each training batch $S_B = \{x_i^s, y_i^s\}_{i=1}^B$ with B samples from

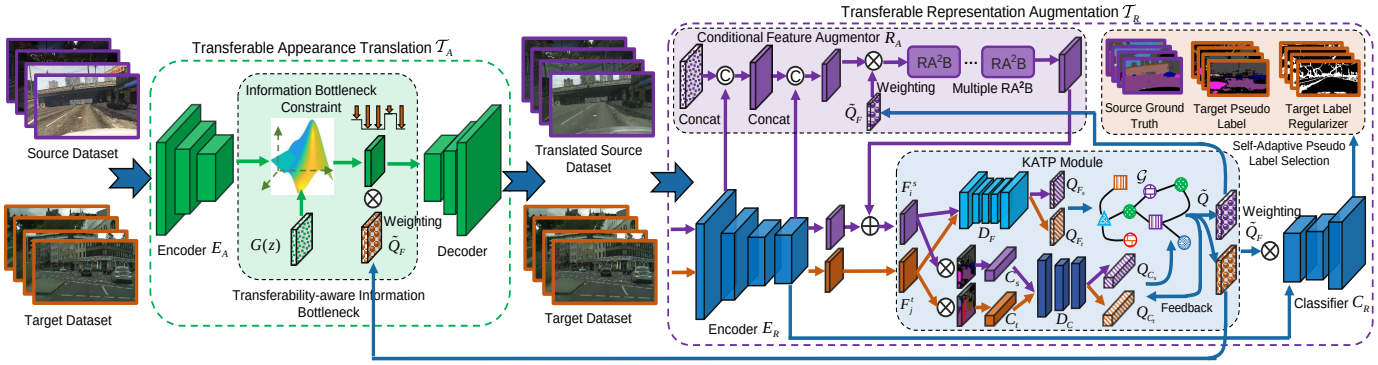


Fig. 2. Graphical illustration of our KATPAN, including *a knowledge aggregation-induced transferability perception* (KATP) module to quantify which semantic representations are transferable or not across domains, *a transferable appearance translation module* $\mathcal{T}_A(\cdot)$ to capture where to highlight transferable visual appearances while abandoning untransferable visual translation, and *a transferable representation augmentation module* $\mathcal{T}_R(\cdot)$ to determine how to augment transferable representations while preventing the negative transfer brought by irrelevant representations.

source domain, the estimated source prototype $\hat{C}_s^k$ of the k-th class is defined as the mean embedding of all source features belonging to the k-th class ($k = 1, \cdots, K$) in S_B:

$$\hat{C}_s^k = \mathbb{E}_{(x_i^s, y_i^s) \in S_B} \big[\frac{1}{N_s^k} \sum_{i=1}^{B} \sum_{m=1}^{|x_i^s|} \big((F_i^s)_m \cdot \mathbf{1}_{(y_i^s)_m = k} \big) \big], \quad (3)$$

where $(F_i^s)_m$ represents the extracted feature at the m-th pixel of the i-th source sample in S_B, $(y_i^s)_m$ denotes its corresponding label, and $N_s^k = \sum_{i=1}^{B} \sum_{m=1}^{|x_i^s|} \mathbf{1}_{(y_i^s)_m = k}$ is the number of pixels belonging to the k-th class in S_B. Unfortunately, the ground truths in target domain are unavailable, and we assign confident pseudo labels to target samples via the self-adaptive pseudo label selection strategy in Section 3.3.2, as presented in Fig. 2. To be specific, in the mini-batch $T_B = \{x_j^t\}_{j=1}^{B}$ of target domain, the estimated target prototype $\hat{C}_t^k$ of the k-th class is formulated as follows:

$$\hat{C}_t^k = \mathbb{E}_{x_j^t \in T_B} \big[\frac{1}{N_t^k} \sum_{j=1}^{B} \sum_{n=1}^{|x_j^t|} \big((F_j^t)_n \cdot \mathbf{1}_{(\hat{y}_j^t)_n = k} \big) \big], \quad (4)$$

where $N_t^k = \sum_{j=1}^{B} \sum_{n=1}^{|x_j^t|} \mathbf{1}_{(\hat{y}_j^t)_n = k}$ indicates the number of pixels with pseudo labels as the k-th class in T_B. $(F_j^t)_n$ and $(\hat{y}_j^t)_n$ denote the extracted representation and the mined pseudo label at the n-th pixel of the j-th target image in T_B. In order to compensate the prototypes estimation bias brought by the sampling randomness of mini-batch, for the k-th class, we construct global category-wise prototypes C_s^k and C_t^k across source and target domains via an exponential moving average strategy:

$$\begin{aligned} C_s^k &= \gamma C_s^k + (1-\gamma)\hat{C}_s^k, \\ C_t^k &= \gamma C_t^k + (1-\gamma)\hat{C}_t^k, \end{aligned} \quad (5)$$

where C_s^k and C_t^k are updated along with the iterative training batches, and γ represents the exponential decay weight that is empirically set as 0.7 in our experiments.

As depicted in Fig. 2, we respectively concatenate global category-wise prototypes from source and target domains as $C_s = [C_s^1, \cdots, C_s^K]^\top \in \mathbb{R}^{K \times U}$ and $C_t = [C_t^1, \cdots, C_t^K]^\top \in \mathbb{R}^{K \times U}$, which are then forwarded into the category-level discriminator $D_C(\cdot)$ to identify which one is from source or target domain. Given the probability outputs $D_C(C_s; \theta_{D_C})$ and $D_C(C_t; \theta_{D_C})$ predicted by discriminator $D_C(\cdot)$, the quantified transferability $Q_{C_s} \in \mathbb{R}^K$ and $Q_{C_t} \in \mathbb{R}^K$ for C_s and C_t can be respectively denoted as follows:

$$\begin{aligned} Q_{C_s} &= 1 - \mathcal{I}(D_C(C_s; \theta_{D_C})), \\ Q_{C_t} &= 1 - \mathcal{I}(D_C(C_t; \theta_{D_C})), \end{aligned} \quad (6)$$

where θ_{D_C} is the network weight of $D_C(\cdot)$.

• **Transferability Knowledge Graph:** To propagate transferability of constructed category-wise prototypes, the representations of these prototypes and pixels are structured as a transferability knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. In this graph, the vertex set $\mathcal{V}$ is composed of source feature $F_s^r \in \mathbb{R}^{(H_s W_s) \times U}$ reshaped from $F_i^s \in \mathbb{R}^{H_s \times W_s \times U}$, target feature $F_t^r \in \mathbb{R}^{(H_t W_t) \times U}$ reshaped from $F_j^t \in \mathbb{R}^{H_t \times W_t \times U}$, source and target prototypes ($C_s \in \mathbb{R}^{K \times U}$, $C_t \in \mathbb{R}^{K \times U}$). We concatenate them together as the representation matrix $\mathcal{M} \in \mathbb{R}^{|\mathcal{V}| \times U}$, where $|\mathcal{V}| = H_s W_s + H_t W_t + 2K$. Similar to $\mathcal{M}$, the transferability corresponding to the features in $\mathcal{M}$ could be concatenated together as $\mathcal{Q} = [Q_{F_s}^r, Q_{F_t}^r, Q_{C_s}, Q_{C_t}] \in \mathbb{R}^{|\mathcal{V}|}$, where $Q_{F_s}^r \in \mathbb{R}^{H_s W_s}$ and $Q_{F_t}^r \in \mathbb{R}^{H_t W_t}$ are respectively reshaped from $Q_{F_s} \in \mathbb{R}^{H_s \times W_s}$ and $Q_{F_t} \in \mathbb{R}^{H_t \times W_t}$. Besides, the edge set $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ characterizes the intrinsic relationships among vertexes, and the adjacency matrix $\mathcal{A} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$ is utilized to model such relations. To be specific, a Gaussian kernel over pairs of vertexes in $\mathcal{V}$ is employed to derive the adjacency matrix $\mathcal{A}$:

$$\mathcal{A}_{m,n} = \exp\Big(-\frac{\|\mathcal{M}_m - \mathcal{M}_n\|_2^2}{2\varrho^2}\Big), \quad (7)$$

where $\mathcal{M}_m \in \mathbb{R}^U$ and $\mathcal{M}_n \in \mathbb{R}^U$ respectively denote the representations at the m-th and n-th rows of $\mathcal{M}$. ϱ is the standard deviation parameter of Gaussian function.

Moreover, we employ the adjacency matrix $\mathcal{A}$ to refine the quantified transferability perception $\mathcal{Q}$, which could capture intrinsic transferability relationships between semantically adjacent representations and global category-wise prototypes. In the refining process, for each pixel, we can aggregate the knowledge from its adjacent area. Consequently, the knowledge aggregation-induced transferability $\tilde{\mathcal{Q}} \in \mathbb{R}^{|\mathcal{V}|}$ is formulated as:

$$\tilde{\mathcal{Q}} = \mathcal{D}^{-\frac{1}{2}} \mathcal{A} \mathcal{D}^{-\frac{1}{2}} \mathcal{Q} = [\tilde{Q}_{F_s}^r, \tilde{Q}_{F_t}^r, \tilde{Q}_{C_s}, \tilde{Q}_{C_t}], \quad (8)$$

where $\mathcal{D} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ is the diagonal degree matrix with entries $\mathcal{D}_{ii} = \sum_j \mathcal{A}_{ij}$, and $\mathcal{D}^{-\frac{1}{2}}$ is used to normalize the adjacency matrix $\mathcal{A}$ [2], [31]. Such normalization strategy promotes the transferability propagation between category-wise prototypes and semantically adjacent features. The $\tilde{Q}^r_{F_s} \in \mathbb{R}^{H_s W_s}$ and $\tilde{Q}^r_{F_t} \in \mathbb{R}^{H_t W_t}$ are respectively reshaped as $\tilde{Q}_{F_s} \in \mathbb{R}^{H_s \times W_s}$ and $\tilde{Q}_{F_t} \in \mathbb{R}^{H_t \times W_t}$, which are both denoted as $\tilde{Q}_F$ for simplification in this paper.

3.3 KATPAN: A KATP Adaptation Network

With the quantified transferability in Eq. (8), we aim to achieve *where to transfer* by weighting information bottleneck constraint in transferable appearance translation $\mathcal{T}_A(\cdot)$, and achieve *how to transfer* by weighting semantic features in transferable representation augmentation $\mathcal{T}_R(\cdot)$.

3.3.1 Where To Transfer: Transferable Appearance Translation $\mathcal{T}_A(\cdot)$

Recently, unpaired image-to-image translation (*a.k.a.*, style transfer [17], [53]) has shown a powerful ability to perform image alignment without using labels [20]. In light of this, we consider proposing a module $\mathcal{T}_A(\cdot)$ to explore transferable translation while neglecting untransferable translation. As shown in Fig. 2, $\mathcal{T}_A(\cdot)$ uses a vanilla appearance translation module [53] (*i.e.*, CycleGAN) as the basic module. Then, to capture where to translate transferable visual characterizations while neglecting untransferable appearance translation across domains, we modify CycleGAN by introducing a transferability-aware information bottleneck.

• **CycleGAN:** To learn the appearance mappings $X_s \rightarrow X_t$ and $X_t \rightarrow X_s$, we forward both X_s and X_t into the appearance translation module CycleGAN in $\mathcal{T}_A(\cdot)$, and it generates the corresponding translated source data $\hat{X}_s = \{\hat{x}^s_i, y^s_i\}_{i=1}^{n_s}$ and translated target data $\hat{X}_t = \{\hat{x}^t_j\}_{j=1}^{n_t}$. $\hat{x}^s_i = \mathcal{T}_A(x^s_i; \theta_{\mathcal{T}_A})$ and $\hat{x}^t_j = \mathcal{T}_A^{-1}(x^t_j; \theta_{\mathcal{T}_A^{-1}})$ are the translated images from $\hat{X}_s$ and $\hat{X}_t$, where $\theta_{\mathcal{T}_A}$ and $\theta_{\mathcal{T}_A^{-1}}$ denote the parameters of $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_A^{-1}(\cdot)$, and $\mathcal{T}_A^{-1}(\cdot)$ indicates the reverse mapping of $\mathcal{T}_A(\cdot)$ (*i.e.*, $X_t \rightarrow \hat{X}_s$). Note that original and translated source images (x^s_i and $\hat{x}^s_i$) share the same pixel annotation y^s_i, in spite of large visual divergence between them. For the mapping $X_s \rightarrow X_t$, $\mathcal{L}_{\text{gan}}(\hat{X}_s, X_t, \mathcal{T}_A, D_1)$ is utilized to mitigate the distribution gap between $\hat{X}_s$ and X_t:

$$\begin{aligned}\mathcal{L}_{\text{gan}}(\hat{X}_s, X_t, \mathcal{T}_A, D_1) = \mathbb{E}_{x^t_j \in X_t}\big[\log(D_1(x^t_j; \theta_{D_1}))\big] \\ + \mathbb{E}_{x^s_i \in X_s}\big[1 - \log(D_1(\mathcal{T}_A(x^s_i; \theta_{\mathcal{T}_A}); \theta_{D_1}))\big],\end{aligned} \tag{9}$$

where $D_1(\cdot)$ denotes the discriminator that distinguishes the real target image x^t_j from the translated source image $\hat{x}^s_i$. θ_{D_1} represents the network parameters of $D_1(\cdot)$. Likewise, $\mathcal{L}_{\text{gan}}(\hat{X}_t, X_s, \mathcal{T}_A^{-1}, D_2)$ is employed to learn the translation $X_t \rightarrow X_s$, where $D_2(\cdot)$ shares similar definition with $D_1(\cdot)$, but identifies whether the input is from the translated target image $\hat{x}^t_j$ or the real source image x^s_i. θ_{D_2} denotes the network weights of $D_2(\cdot)$. Furthermore, we utilize $\mathcal{L}_{\text{cyc}}(X_s, X_t, \mathcal{T}_A, \mathcal{T}_A^{-1})$ to encourage semantic consistency between the real and translated images across domains:

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(X_s, X_t, \mathcal{T}_A, \mathcal{T}_A^{-1}) = \mathbb{E}_{x^s_i \in X_s}\big[\big\|\mathcal{T}_A^{-1}(\hat{x}^s_i; \theta_{\mathcal{T}_A^{-1}}) - x^s_i\big\|_1\big] \\ + \mathbb{E}_{x^t_j \in X_t}\big[\big\|\mathcal{T}_A(\hat{x}^t_j; \theta_{\mathcal{T}_A}) - x^t_j\big\|_1\big].\end{aligned} \tag{10}$$

Therefore, the training objective $\mathcal{L}_{\mathcal{T}_A}$ is summarized as:

$$\begin{aligned}\mathcal{L}_{\mathcal{T}_A} =& \mathcal{L}_{\text{gan}}(\hat{X}_s, X_t, \mathcal{T}_A, D_1) + \mathcal{L}_{\text{gan}}(\hat{X}_t, X_s, \mathcal{T}_A^{-1}, D_2) \\ &+ \alpha \mathcal{L}_{\text{cyc}}(X_s, X_t, \mathcal{T}_A, \mathcal{T}_A^{-1}),\end{aligned} \tag{11}$$

where α is the balanced weight. However, Eq. (11) has no effective constraint to prevent $\mathcal{T}_A(\cdot)$ from encoding adaptation-independent nuisance factors (*i.e.*, untransferable knowledge) into the latent features. Moreover, it cannot capture transferable appearance translation with high transferability while abandoning untransferable translation.

• **Transferability-aware Information Bottleneck:** To tackle the above challenges, as depicted in Fig. 2, we develop a transferability-aware information bottleneck to capture transferable appearance translation. We first consider purifying adaptation-dependent semantic representations (*i.e.*, transferable knowledge), and then weight them with different transferability.

1. To purify adaptation-dependent latent representations (*i.e.*, transferable knowledge), relying on the information theory [1], [34], we integrate an information bottleneck constraint on the latent feature space. [1], [34] have experimentally shown that the latent representations will be more discriminative[1] if they are closer to a Gaussian distribution. Thus, we use a distributional discrepancy measure $\text{Dist}(\cdot,\cdot)$ between a Gaussian distribution $G(z)$ and latent representations (extracted by the $E_A(\cdot)$ in $\mathcal{T}_A(\cdot)$) to filter non-discriminative latent representations by setting a threshold T_b. To be specific, Eq. (11) is reformulated as follows:

$$\begin{aligned}\mathcal{L}_{\mathcal{T}_A} = \mathcal{L}_{\text{gan}}(\hat{X}_s, X_t, \mathcal{T}_A, D_1) + \mathcal{L}_{\text{gan}}(\hat{X}_t, X_s, \mathcal{T}_A^{-1}, D_2) \\ + \alpha \mathcal{L}_{\text{cyc}}(X_s, X_t, \mathcal{T}_A, \mathcal{T}_A^{-1}), \\ s.t.\ \mathbb{E}_{x^s_i \in X_s}\big[\big\|\text{Dist}(E_A(x^s_i; \theta_{E_A}), G(z))\big\|_{1,1}\big] \le T_b, \\ \mathbb{E}_{x^t_j \in X_t}\big[\big\|\text{Dist}(E_A(x^t_j; \theta_{E_A}), G(z))\big\|_{1,1}\big] \le T_b,\end{aligned} \tag{12}$$

where T_b is a scalar threshold, and θ_{E_A} represents the weights of $E_A(\cdot)$. Following [1], [34], we use the Kullback-Leibler (KL) divergence as $\text{Dist}(\cdot,\cdot)$ and set $G(z)$ as a standard Gaussian $\mathcal{N}(0; I)$ in this paper. $\|\cdot\|_{1,1}$ is the $1,1$-norm for matrix (*i.e.*, $\|A\|_{1,1} = \frac{1}{mn}\sum_{i=1}^m \sum_{j=1}^n |A_{ij}|$, here A is a $m \times n$ matrix).

Based on the information bottleneck constraints in Eq. (12), the above optimization problem focuses on minimizing $\mathcal{L}_{\mathcal{T}_A}$ by exploring discriminative features across source and target domains while preventing $\mathcal{T}_A(\cdot)$ from encoding non-discriminative features. Namely, we will find discriminative and adaptation-dependent latent representations across domains (*i.e.*, transferable knowledge) when this optimization problem is solved. It can be also seen that the non-discriminative features and adaptation-independent nuisance factors (*i.e.*, untransferable knowledge) of both domains are discarded when minimizing $\mathcal{L}_{\mathcal{T}_A}$. Therefore, when performing unpaired image-to-image translation via minimizing Eq. (12), the information bottleneck constraint in Eq. (12) encourages $\mathcal{T}_A(\cdot)$ to neglect untransferable appearance translation while exploring where to translate transferable appearance. To solve the optimization

1. Here, discriminative features represent the features that can be used to well distinguish samples rather than discriminating domains.

problem in Eq. (12), we introduce two Lagrange multipliers $\lambda_s \geq 0$ and $\lambda_t \geq 0$ into Eq. (12), and equally rewrite it as:

$$\begin{aligned}\mathcal{L}_{\mathcal{T}_A} =&\mathcal{L}_{\rm gan}(\hat{X}_s, X_t, \mathcal{T}_A, D_1) + \mathcal{L}_{\rm gan}(\hat{X}_t, X_s, \mathcal{T}_A^{-1}, D_2)\\ &+\alpha\mathcal{L}_{\rm cyc}(X_s, X_t, \mathcal{T}_A, \mathcal{T}_A^{-1})\\ &+\lambda_s\big(\mathbb{E}_{x_i^s\in X_s}\big[\big\|{\rm Dist}(E_A(x_i^s;\theta_{E_A}), G(z))\big\|_{1,1} - T_b\big]\big)\\ &+\lambda_t\big(\mathbb{E}_{x_j^t\in X_t}\big[\big\|{\rm Dist}(E_A(x_j^t;\theta_{E_A}), G(z))\big\|_{1,1} - T_b\big]\big).\end{aligned} \tag{13}$$

We denote the last two terms in Eq. (13) as $\mathcal{L}_b^s$ and $\mathcal{L}_b^t$, *i.e.*, $\mathcal{L}_b^s = \mathbb{E}_{x_i^s\in X_s}\big[\big\|{\rm Dist}(E_A(x_i^s;\theta_{E_A}), G(z))\big\|_{1,1} - T_b\big]$ and $\mathcal{L}_b^t = \mathbb{E}_{x_j^t\in X_t}\big[\big\|{\rm Dist}(E_A(x_j^t;\theta_{E_A}), G(z))\big\|_{1,1} - T_b\big]$. $\mathcal{L}_b^s$ and $\mathcal{L}_b^t$ are employed to adaptively update λ_s and λ_t via dual gradient descent. The intuitive illustration behind is that performing a specific information bottleneck constraint on latent space is critical to adaptively purify adaptation-dependent knowledge (*i.e.*, transferable knowledge). To this end, after initializing λ_s and λ_t, we adaptively update λ_s and λ_t via the network itself along the iteration optimization process, *i.e.*, $\lambda_s \leftarrow \max(0, \lambda_s + \nu\mathcal{L}_b^s)$ and $\lambda_t \leftarrow \max(0, \lambda_t + \nu\mathcal{L}_b^t)$. $\nu = \lambda_s/\lambda_t$ denotes the step length to update λ_s and λ_t.

2. Although the information bottleneck constraint in Eq. (13) effectively removes adaptation-independent information (*i.e.*, untransferable knowledge), it neglects the different transferability contributions of adaptation-dependent latent representations (*i.e.*, transferable knowledge) from different classes. Moreover, the objects belonging to the same class but with different appearances and modalities, have unequal contributions to transferable appearance translation. To address these limitations in Eq. (13), we weight the adaptation-dependent semantic representations (*i.e.*, transferable knowledge) using the quantified transferability perception $\tilde{Q}_F$ in Eq. (8). Specifically, the information bottleneck constraint (*i.e.*, KL divergence) in Eq. (13) is adaptively weighted by multiplying $\tilde{Q}_F$, and we reformulate Eq. (13) as follows:

$$\begin{aligned}\mathcal{L}_{\mathcal{T}_A} =& \mathcal{L}_{\rm gan}(\hat{X}_s, X_t, \mathcal{T}_A, D_1) + \mathcal{L}_{\rm gan}(\hat{X}_t, X_s, \mathcal{T}_A^{-1}, D_2)\\ &+ \alpha\mathcal{L}_{\rm cyc}(X_s, X_t, \mathcal{T}_A, \mathcal{T}_A^{-1})\\ &+ \lambda_s\big(\mathbb{E}_{x_i^s\in X_s}\big[\big\|\tilde{Q}_F \odot {\rm Dist}(E_A(x_i^s;\theta_{E_A}), G(z))\big\|_{1,1} - T_b\big]\big)\\ &+ \lambda_t\big(\mathbb{E}_{x_j^t\in X_t}\big[\big\|\tilde{Q}_F \odot {\rm Dist}(E_A(x_j^t;\theta_{E_A}), G(z))\big\|_{1,1} - T_b\big]\big),\end{aligned} \tag{14}$$

where $\odot$ represents the Hadamard product. Intuitively, the adaptation-dependent semantic information with high transfer score $\tilde{Q}_F$ will get stronger information constraint to explore more transferable appearance translations by multiplying $\tilde{Q}_F$, and vise vice. Such transferability weighting strategy in Eq. (14) considers different transferability contributions of adaptation-dependent information to transferable appearance translation, and thus achieves where to transfer transferable visual translation. It increases the transferability attention on important transferable information, while preventing the critical information from being eliminated.

3.3.2 How To Transfer: Transferable Representation Augmentation $\mathcal{T}_R(\cdot)$

Even though $\mathcal{T}_A(\cdot)$ highlights where to translate transferable visual appearances, it cannot encourage the distribution discrepancy across domains to be mitigated effectively. To this end, transferable representation augmentation module $\mathcal{T}_R(\cdot)$ is designed to explore how to augment transferable representations, which further minimizes domain gap and improves the appearance adaptation performance of $\mathcal{T}_A(\cdot)$ in return. As shown in Fig. 2, $\mathcal{T}_R(\cdot)$ consists of a conditional feature augmentor $R_A(\cdot)$, a self-adaptive pseudo label selection strategy and a category-level adversarial framework.

• **Conditional Feature Augmentor:** With the quantified transferability $\tilde{Q}_F$ in Eq. (8), conditional feature augmentor $R_A(\cdot)$ could effectively explore how to augment the transferable knowledge while neglecting the irrelevant feature augmentation. Generally, $R_A(\cdot)$ focuses on capturing the residual noisy feature between the source and target representations to characterize domain discrepancy, and compensates the domain discrepancy by performing transferable feature augmentation on source features.

1. Feature Augmentation: To be specific, as presented in Fig. 2, $R_A(\cdot)$ takes a noise vector from Gaussian distribution $G(z) = \mathcal{N}(0; I)$ as input, and then concatenates the noisy input, low-level and high-level source representations extracted from feature extractor $E_R(\cdot)$. The concatenated feature is then weighted by the transferability perception $\tilde{Q}_F$ in Eq. (8), before we forward it into multiple residual attention on attention blocks ($\rm RA^2B$) to obtain the residual noisy feature. For the i-th sample $\hat{x}_i^s$ from translated source dataset $\hat{X}_s$, we denote the source representation extracted by $E_R(\cdot)$ as $E_R(\hat{x}_i^s;\theta_{E_R})$, and the residual noisy feature of $R_A(\cdot)$ as $R_A(G(z), \hat{x}_i^s;\theta_{R_A})$, where θ_{R_A} and θ_{E_R} denotes the parameters of $R_A(\cdot)$ and $E_R(\cdot)$. We replace the original source representation with the augmented source feature $F_i^s = E_R(\hat{x}_i^s;\theta_{E_R}) + R_A(G(z), \hat{x}_i^s;\theta_{R_A})$ to perform transferable feature augmentation, since we intuitively observe that the residual noisy feature could effectively compensate the distribution discrepancy across domains. The augmented source feature F_i^s is encouraged to preserve the semantic information of $\hat{x}_i^s$, and appears as if it is extracted from the target domain simultaneously. Thus, both augmented source feature F_i^s and target feature $F_j^t = E_R(x_j^t;\theta_{E_R})$ extracted via $E_R(\cdot)$ for the j-th target sample are fed into the discriminator $D_F(\cdot)$, to identify which one comes from source or target domain. The objective is formally presented as follows:

$$\begin{aligned}\min_{\theta_{R_A},\theta_{E_R}}\max_{\theta_{D_F}}\mathcal{L}_{D_F} =&\mathbb{E}_{x_j^t\in X_t}[\log(1 - D_F(F_j^t;\theta_{D_F})]\\ &+\mathbb{E}_{\hat{x}_i^s\in\hat{X}_s}[\log(D_F(F_i^s;\theta_{D_F}))],\end{aligned} \tag{15}$$

where θ_{D_F} represents the network parameters of $D_F(\cdot)$.

Intuitively, when optimizing the objective Eq. (15), the transferability perception $\tilde{Q}_F$ that weights the concatenated features in $R_A(\cdot)$ would encourage $R_A(\cdot)$ to explore how to augment transferable source representations. Thus, $R_A(\cdot)$ could capture more transferable feature augmentation while neglecting untransferable augmentation. As presented in Fig. 2, the probability output of $D_F(\cdot)$ is employed to quantify the knowledge aggregation-induced transferability $\tilde{Q}_F$, which is then forwarded into $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$, respectively.

2. Residual Attention on Attention Block ($\rm RA^2B$): In $R_A(\cdot)$, $\rm RA^2B$ focuses on encoding residual noisy feature to perform transferable representation augmentation on source feature. It also captures the transferability contribution of source transferable knowledge, and transmits it into residual noisy feature. The details of $\rm RA^2B$ are shown as follows.

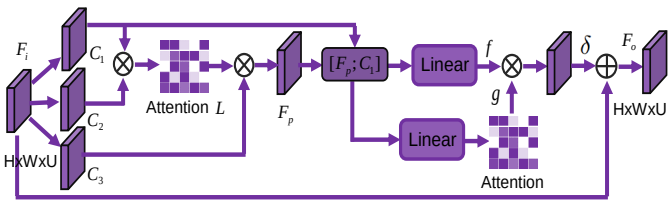


Fig. 3. Graphical demonstration of the $\mathrm{RA}^2\mathrm{B}$ module.

As presented in Fig. 3, the input feature $F_i \in \mathbb{R}^{H\times W\times U}$ is fed into three convolution layers to respectively generate three new representations C_1, C_2 and C_3 ($C_1, C_2, C_3 \in \mathbb{R}^{H\times W\times U}$), where H, W and U are the height, width and channel dimensions. We then reshape C_1 and C_2 as $\mathbb{R}^{N\times U}$ ($N = H \times W$), and obtain the attention $L = \{L_{ij}\}_{i,j=1}^{N} \in \mathbb{R}^{N\times N}$ via softmax activation, where $L_{ij} = \exp((C_1 C_2^\top)_{ij}) / \sum_{i=1}^{N} \exp((C_1 C_2^\top)_{ij})$. The matrix multiplication among the transpose of L and the reshaped $C_3 \in \mathbb{R}^{N\times U}$ is used to obtain the attention feature $F_p \in \mathbb{R}^{N\times U}$, where $(F_p)_i = \sum_{j=1}^{N} L_{ij}(C_3)_j$ and $(C_3)_j$ are the features at the i-th pixel of F_p and the j-th pixel of C_3, respectively.

Unfortunately, the above attention mechanism still produces a weighted average feature F_p, even though there are no any transferable representations. It is unknown that whether or how well the attention result is related to the transferable knowledge. Moreover, it cannot filter out fallacious attention results and only keep the useful ones. When obtaining false attention results, it could prevent $R_A(\cdot)$ from exploring transferable representation augmentation by misjudging transferable features as untransferable knowledge. Thus, $\mathrm{RA}^2\mathrm{B}$ is designed to address fallacious attention results, and capture the transferability contribution of source transferable knowledge by further evaluating the semantic relevance between F_p and C_1. To be specific, the linear operation between F_p and C_1 is employed to generate the transferable information flow f and the relevance gate g:

$$\begin{aligned} f &= W_f^1 C_1 + W_f^2 F_p + B_f, \\ g &= \sigma(W_g^1 C_1 + W_g^2 F_p + B_g), \end{aligned} \tag{16}$$

where $W_f^1, W_f^2, W_g^1, W_g^2 \in \mathbb{R}^{N\times N}$, $B_f, B_g \in \mathbb{R}^{N\times U}$ are the transformation matrices. σ is the sigmoid activation function. Then, we reshape f and g as $\mathbb{R}^{H\times W\times U}$, and perform element-wise multiplication among them. The output matrix is multiplied by a scalar δ, and further performs an element-wise sum operation with F_i to produce the ultimate representation $F_o = \delta \cdot (f \odot g) + F_i$, where $\odot$ indicates the Hadamard product. We initialize δ as 0, and optimize it adaptively via our model.

• **Self-adaptive Pseudo Label Selection:** To optimize segmentation loss $\mathcal{L}_{\text{seg}}$, we feed both the translated source example $\hat{x}_i^s$ with pixel label y_i^s and target sample x_j^t with mined pseudo label $\hat{y}_j^t$ into the segmentation network $S(\cdot) = C_R \circ E_R(\cdot)$, where $E_R(\cdot)$ and $C_R(\cdot)$ are the encoder and pixel classifier, and θ_S is the parameters of $E_R(\cdot)$ and $C_R(\cdot)$. As shown in Fig. 2, after extracting the semantic features via $E_R(\cdot)$, we first use transferability perception $\tilde{Q}_F$ in Eq. (8) to weight them, and then forward the transferability-weighted features into classifier $C_R(\cdot)$ for prediction. When optimizing $\mathcal{L}_{\text{seg}}$, such transferability weighting strategy encourages $S(\cdot)$ to pay more attention on transferable features while neglecting untransferable features. $\mathcal{L}_{\text{seg}}$ is formally expressed as follows:

$$\begin{aligned} \min_{\theta_{R_A},\theta_S} \mathcal{L}_{\text{seg}} =& \mathbb{E}_{(\hat{x}_i^s, y_i^s)\in \hat{X}_s}\big[-\sum_{m=1}^{|\hat{x}_i^s|}\sum_{k=1}^{K} \mathbf{1}_{(y_i^s)_m=k} \log(S(\hat{x}_i^s;\theta_S)_m^k)\big] \\ &+ \mathbb{E}_{(x_j^t,\hat{y}_j^t)\in X_t}\big[-\sum_{n=1}^{|x_j^t|}\sum_{k=1}^{K}\Big(\frac{\mathbf{1}_{(\hat{y}_j^t)_n=k}\log(S(x_j^t;\theta_S)_n^k)}{\vartheta_j^k} \\ &+ \mathcal{R}_P(S(x_j^t;\theta_S)_n, (\hat{y}_j^t)_n) + \mathcal{R}_E(S(x_j^t;\theta_S)_n, (\hat{y}_j^t)_n)\Big)\big], \\ &s.t.\ \ (\hat{y}_j^t)_n \subset \{\mathbf{e}_k \cup \mathbf{0} | \mathbf{e}_k \in \mathbb{R}^K\}, \end{aligned} \tag{17}$$

where $S(\hat{x}_i^s;\theta_S)_m^k$ and $S(x_j^t;\theta_S))_n^k$ respectively indicate the output probability of $S(\cdot)$ predicted as the k-th class at the m-th and the n-th pixels. $\{\vartheta_j^k\}_{k=1}^K$ denotes the self-adaptive selection threshold to mine one-hot or all-zero pseudo label $\hat{y}_j^t$, where the n-th pixel position of $\hat{y}_j^t$ is determined by:

$$(\hat{y}_j^t)_n^k = \begin{cases} 1, \text{if } \ k = \arg\max_k S(x_j^t;\theta_S)_n \text{ and } S(x_j^t;\theta_S)_n > \vartheta_j^k \\ 0, \qquad\qquad \text{otherwise.} \end{cases} \tag{18}$$

To adaptively determine the selection of $\{\vartheta_j^k\}_{k=1}^K$, different from [54] utilizing all target samples, we sort the probability outputs of the j-th target sample that are predicted as the k-th class under the descending order, which is denoted as $\mathcal{P}_j^k$ for simplification. ϑ_j^k can then be determined by:

$$\vartheta_j^k = \gamma\vartheta_j^k + (1-\gamma)\mathcal{P}_j^k\big[\kappa(\vartheta_j^k)^\tau |\mathcal{P}_j^k|\big], \tag{19}$$

where $\kappa \in [0,1]$ controls the selection portion of pseudo labels, and $(\vartheta_j^k)^\tau$ adaptively modifies the selection portion to prevent the generation of noise labels. $\gamma = 0.7$ and $\tau = 8$ denote the weight decay. $\{\vartheta_j^k\}_{k=1}^K$ are initialized as 0.9.

Moreover, $\mathcal{R}_P(S(x_j^t;\theta_S)_n, (\hat{y}_j^t)_n) = \mathcal{Z}_n \cdot S(x_j^t;\theta_S)_n / K$ represents the label regularizer to smooth the confident prediction of pseudo labels and further restrain the noise generation, where $\mathcal{Z}_n = \mathbf{1}_{(\hat{y}_j^t)_n \neq \mathbf{0}}$ is the regularizer mask. Besides, $\mathcal{R}_E(S(x_j^t;\theta_S)_n, (\hat{y}_j^t)_n) = (1-\mathcal{Z}_n)\cdot S(x_j^t;\theta_S)_n \cdot \log\big(S(x_j^t;\theta_S)_n\big)$ is the entropy regularizer to minimize the uncertainty of unconfident prediction and encourage the confident pseudo label prediction to be sharper.

• **Category-level Adversarial Framework:** The updated source and target global category-wise prototypes (*i.e.*, $C_s = [C_s^1, \cdots, C_s^K]^\top$ and $C_t = [C_t^1, \cdots, C_t^K]^\top$) are fed into the category-level discriminator $D_C(\cdot)$ to further mitigate the domain discrepancy, and the category-level adversarial objective is concretely formulated as follows:

$$\begin{aligned} \min_{\theta_{R_A},\theta_{E_R}} \max_{\theta_{D_C}} \mathcal{L}_{D_C} =& \mathbb{E}_{x_j^t\in X_t}[\log(1 - D_C(C_t;\theta_{D_C})] \\ &+ \mathbb{E}_{x_i^s\in X_s}[\log(D_C(C_s;\theta_{D_C}))]. \end{aligned} \tag{20}$$

In Eq. (20), each category is paid equal attention. However, in the real world, different categories have different contributions for domain adaptation (see Section 3.2). To this end, we employ the quantified transferability (*i.e.*, $\tilde{Q}_{C_s}$ and $\tilde{Q}_{C_t}$) in Eq. (8) to modulate Eq. (20) as follows:

$$\begin{aligned} \min_{\theta_{R_A},\theta_{E_R}} \max_{\theta_{D_C}} \mathcal{L}_{D_C} =& \mathbb{E}_{x_j^t\in X_t}[(\tilde{Q}_{C_t})^\tau \odot \log(1 - D_C(C_t;\theta_{D_C})] \\ &+ \mathbb{E}_{x_i^s\in X_s}[(\tilde{Q}_{C_s})^\tau \odot \log(D_C(C_s;\theta_{D_C}))], \end{aligned} \tag{21}$$

Algorithm 1 : The Optimization of Our KATPAN.

Input: The datasets $\{X_s, X_t\}$, parameters $\{\alpha, T_b, \xi_1, \xi_2\}$;
Output: $\theta_{\mathcal{T}_A}, \theta_{\mathcal{T}_R}$

while $\mathcal{T}_A(\cdot)$ is not converged **do**
  Sample a mini-batch from X_s and X_t;
  Build transferability-aware information bottleneck;
  Optimize $\mathcal{T}_A(\cdot)$ via minimizing Eq. (14);
end while
Obtain the translated datasets $\hat{X}_s$ and $\hat{X}_t$;
while $\mathcal{T}_R(\cdot)$ is not converged **do**
  Sample a mini-batch from $\hat{X}_s$ and X_t;
  Update global category-wise prototypes;
  Obtain initial quantified transferability via optimizing conditional feature augmentor and category-level adversarial framework (*i.e.*, Eq. (15) and Eq. (21));
  Update the quantified transferability via structuring transferability knowledge graph $\mathcal{G}$;
  Obtain confident pseudo labels via the self-adaptive pseudo label selection strategy (*i.e.*, Eq. (18));
  Optimize $\mathcal{T}_R(\cdot)$ via minimizing Eq. (22);
end while
return $\theta_{\mathcal{T}_A}, \theta_{\mathcal{T}_R}$.

where $\tau = 8$ controls the weight on hard-to-adapt global prototypes, and $\odot$ represents the Hadamard product. Intuitively, Eq. (21) is designed to assign more transferability weight on hard-to-adapt global prototypes across domains than the easy-to-adapt classes during the training phase. It could effectively neglect easy-to-adapt examples across domains while focusing on hard-to-adapt representations.

In summary, the overall objective for training $\mathcal{T}_{\mathcal{R}}(\cdot)$ is:

$$\min_{\theta_{R_A},\theta_S} \max_{\theta_{D_F},\theta_{D_C}} \mathcal{L}_{\mathcal{T}_R} = \mathcal{L}_{\text{seg}} + \xi_1 \mathcal{L}_{D_F} + \xi_2 \mathcal{L}_{D_C}, \tag{22}$$

where ξ_1 and ξ_2 represent the balanced weights.

4 Experiments

4.1 Datasets and Evaluation Metric

Medical Endoscopic Dataset [9] is collected from more than 1100 volunteers whose gasteroscope and enteroscopy diagnoses have various lesions, *i.e.*, gastritis, ulcer, cancer, bleeding and polyp. It is composed of 2969 gasteroscope examples and 690 enteroscopy samples, where all gasteroscope images with pixel labels and 300 enteroscopy examples without accessing pixel annotations are respectively considered as source and target domains in the training phase. We use the rest of enteroscopy data for prediction.

Cityscapes [8] is a real-world street scenes dataset that is collected from fifty European cities and finely annotated with 34 categories. It consists of a training subset with 2993 examples, a validation subset with 503 images and a testing subset with 1531 samples.

GTA [35] with 24996 samples is synthesized from fictional city named Los Santos in the computer game Grand Theft Auto V. It has 19 compatible object categories with [8].

SYNTHIA [37] is a large-scale virtual scene dataset, whose the subset called SYNTHIA-RANDCITYSCAPES with 9400 examples is employed for domain adaptation.

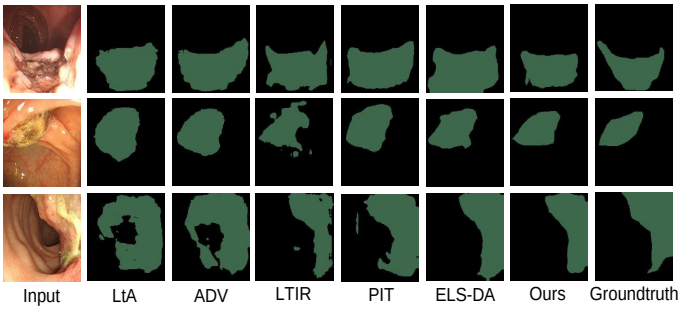


Fig. 4. Some qualitative results on medical endoscopic dataset [9].

NTHU [5] is composed of 4 real-world urban scenes datasets, including Rio, Rome, Taipei and Tokyo, where each dataset consists of the training and testing subsets, and is finely annotated with 13 different object categories.

Evaluation Metric: In our experiments, we utilize Intersection over Union (IoU) as basic metric, and employ three derived metrics for performance evaluation, *i.e.*, mean IoU (mIoU), IoU of disease (IoU_d) and IoU of normal (IoU_n).

4.2 Implementation Details

This subsection introduces the implementation details about network architecture, training and test procedures.

Network Architecture: We utilize CycleGAN [53] as backbone for the transferable appearance translation module $\mathcal{T}_A(\cdot)$. The transferability-aware information bottleneck is attached to the last convolution layer of encoder $E_A(\cdot)$ in $\mathcal{T}_A(\cdot)$, as presented in Fig. 2. In the transferable representation augmentation module $\mathcal{T}_R(\cdot)$, the DeepLab-v2 [3] with ResNet-101 [19] and FCN-8s [38] with VGG-16 [40] are employed as our baseline segmentation network $S(\cdot) = C_R \circ E_R(\cdot)$. Besides, source features from the bottom and top convolution layers of $E_R(\cdot)$ are fed into $R_A(\cdot)$, where the number of RA^2B is set as 16. The discriminator $D_F(\cdot)$ is composed of 5 convolution blocks, while $D_C(\cdot)$ consists of 3 fully-connected layers. We activate each layer of discriminators via the leaky RELU function excluding the last layer activated via the sigmoid function.

Training and Test Stages: In the training phase, two complementary modules $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$ (*i.e.*, Eq. (14) and Eq. (22)) are alternatively optimized until model convergence. Specifically, motivated by [53], we initialize the learning rate as 2.5×10^{-4} in the first 10 epochs and decrease it to 0 linearly in the later 5 epochs when training $\mathcal{T}_A(\cdot)$. In Eq. (14), we set $\alpha = 10$, $T_b = 200$, and initialize both λ_s and λ_t as 1.0×10^{-4}. Moreover, in $\mathcal{T}_R(\cdot)$, we utilize SGD as optimizer for DeepLab-v2 with ResNet-101, where the learning rate is initialized as 2.5×10^{-4} and decreases via poly learning policy. Besides, we use Adam optimizer for FCN-8s with VGG-16, where the initial learning rate is 1.0×10^{-4}, and the momentum of Adam optimizer is 0.9 and 0.99. For $D_F(\cdot)$ and $D_C(\cdot)$, Adam is utilized as the optimizer whose initial learning rate is 1.5×10^{-4} and 1.5×10^{-6} for ResNet-101 and VGG-16, respectively. The optimization of our proposed KATPAN is presented in **Algorithm 1**. In the test phase, the target sample $x_j^t \in X_t$ is directly forwarded into the segmentation network $S(\cdot)$ in $\mathcal{T}_R(\cdot)$ for evaluation.

TABLE 1
Performance comparison between our proposed KATPAN and other competing methods on medical endoscopic dataset [9].

Metrics	Net	Source only	LtA [42]	ADV [43]	BDL [26]	SWES [9]	PyCDA [27]	FADA [44]	CSCL [10]	SS-UDA [33]	PIT [32]	LTIR [23]	ELS-DA [12]	Ours
IoU_n(%)	VGG	67.38	72.75	73.29	75.82	75.10	76.27	77.04	76.61	76.18	77.42	77.08	77.61	**79.34**
IoU_d(%)		25.16	33.40	34.17	36.54	34.86	36.92	37.85	37.28	37.06	38.15	37.75	38.43	**40.27**
mIoU(%)		46.27	53.08	53.73	56.18	54.98	56.60	57.45	56.95	56.62	57.79	57.42	58.02	**59.81**
IoU_n(%)	ResNet	74.47	81.04	81.95	84.22	83.96	83.23	85.23	84.97	82.83	85.62	85.19	85.48	**87.61**
IoU_d(%)		32.65	40.35	42.27	42.84	42.63	42.11	43.48	43.11	40.92	43.88	43.46	43.67	**45.84**
mIoU(%)		53.56	60.70	62.11	63.53	63.29	62.67	64.36	64.04	61.88	64.75	64.33	64.58	**66.73**

TABLE 2
Ablation investigations of our model on medical endoscopic dataset [9].

Variants	Net	KT	TB	PL	AA	mIoU	△(%)
Ours-w/oKT	VGG	✗	✓	✓	✓	56.17	↓3.64
Ours-w/oTB		✓	✗	✓	✓	57.10	↓2.71
Ours-w/oPL		✓	✓	✗	✓	56.45	↓3.36
Ours-w/oAA		✓	✓	✓	✗	58.16	↓1.65
Ours		✓	✓	✓	✓	**59.81**	-
Ours-w/oKT	ResNet	✗	✓	✓	✓	62.81	↓3.92
Ours-w/oTB		✓	✗	✓	✓	64.32	↓2.41
Ours-w/oPL		✓	✓	✗	✓	63.14	↓3.59
Ours-w/oAA		✓	✓	✓	✗	65.25	↓1.48
Ours		✓	✓	✓	✓	**66.73**	-

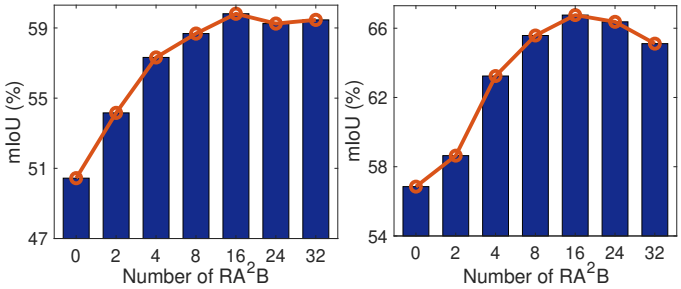


Fig. 5. Effect investigation of RA^2B of our model with VGG (left) and ResNet (right) as backbone on medical endoscopic dataset [9].

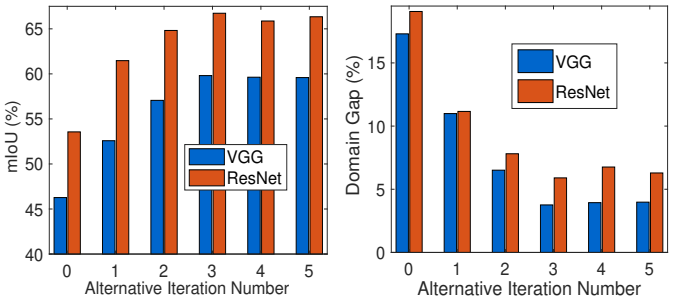


Fig. 6. Effect Investigation of $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$ in terms of mIoU (left) and domain gap (right) on medical endoscopic dataset [9].

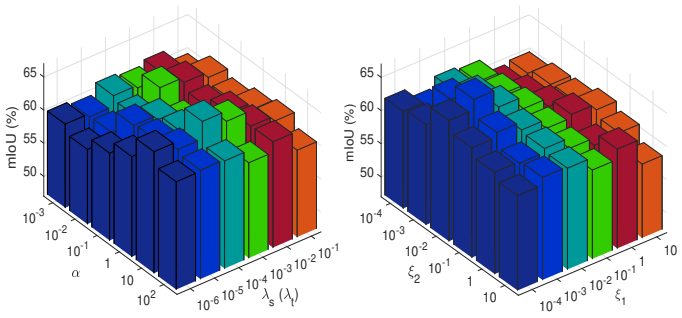
4.3 Experiments on Medical Endoscopic Dataset

4.3.1 Performance Comparison

As shown in Table 1, we investigate the effectiveness of our KATPAN in this subsection, by conducting empirical comparisons between our model and other competing methods on medical endoscopic dataset [9]. We have some essential observations from the performance in Table 1: 1) When compared with the conference version (ELS-DA [12]), our proposed model achieves a large improvement (*i.e.*, $1.79\% \sim 2.15\%$ in terms of mIoU) regardless of different backbone networks. It validates that the knowledge aggregation-induced transferability perception module could effectively achieve transferability propagation between semantically adjacent representations and category-wise prototypes. 2) Different from the previous version (ELS-DA [12]) and BDL [26], our two complementary modules ($\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$) could highlight where and how to capture transferable knowledge across domains by incorporating with the knowledge aggregation-induced transferability perception. 3) The confident pseudo labels for target examples could be mined effectively via our self-adaptive pseudo label selection strategy to minimize the distribution gap across domains. 4) The performance of our model significantly outperforms other competing methods (*e.g.*, [10], [23], [32], [33], [42]) about $1.98\% \sim 6.73\%$ in terms of mIoU, which demonstrates the effectiveness of our model to capture transferable representations while neglecting untransferable knowledge. In addition, some qualitative results in Fig. 4 also support the superior performance of the proposed KATPAN model.

4.3.2 Ablation Study

This subsection reports the ablation investigations of different modules in our proposed model via extensive variant experiments on medical endoscopic dataset [9]. As presented in Table 2, we denote the different components of our model, *i.e.*, knowledge aggregation-induced transferability perception, transferability-aware information bottleneck, self-adaptive pseudo label selection and residual attention on attention of RA^2B as KT, TB, PL and AA, respectively. Moreover, we denote our proposed model trained without KT, TB, PL or AA as Ours-w/oKT, Ours-w/oTB, Ours-w/oPL and Ours-w/oAA in our experiments. Note that Ours-w/oKT also does not use the quantified transferability perception in our previous conference version [12] to perform ablation studies. When any one of modules in our proposed model is abandoned, the adaptation performance of our model degrades $1.48\% \sim 3.64\%$ in terms of mIoU. It illustrates the effectiveness and importance of each component in our model to cooperatively achieve the best transfer performance. Specifically, our proposed KATPAN with the knowledge aggregation-induced transferability perception module is more effective to capture transferable representations across domains and prevent the irrelevant knowledge transfer, when compared with the conference version [12]. Furthermore, the performance of our model decreases $1.48\% \sim 1.65\%$ mIoU after removing RA^2B. It illustrates that the semantic transferability of transferable representations can be highlighted via multiple RA^2Bs. We set the number of RA^2B as 16 according to Fig. 5. Moreover, confident pseudo labels of target samples mined via our self-adaptive pseudo label selection strategy could effectively narrow the distribution discrepancy across domains.



(a) $\xi_1 = 10^{-3}, \xi_2 = 10^{-2}$ (b) $\alpha = 10, \lambda_s(\lambda_t) = 10^{-4}$

Fig. 7. Parameter sensitivity investigation of $\{\alpha, \lambda_s(\lambda_t)\}$ (a) and $\{\xi_1, \xi_2\}$ (b) in terms of mIoU on medical endoscopic dataset [9].

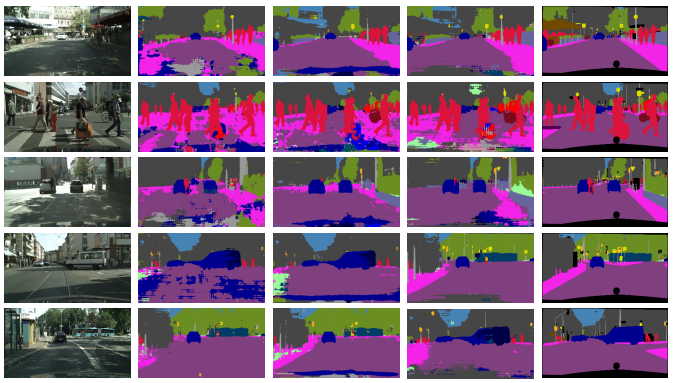


Fig. 8. Some qualitative results on GTA [35] → Cityscapes [8] task.

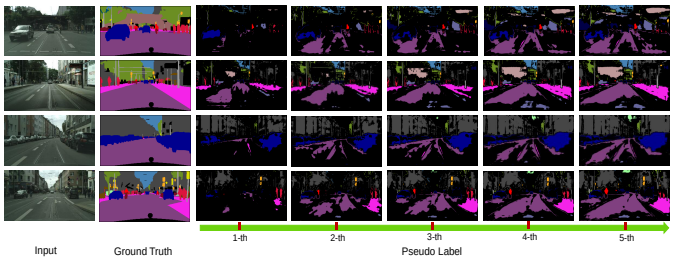


Fig. 9. Pseudo labels visualization on GTA [35] → Cityscapes [8] task.

TABLE 3
Investigation of p value of t-test on medical endoscopic dataset [9].

Variants	VGG	ResNet
Ours vs SWES [9]	1.81×10^{-4}	6.28×10^{-4}
Ours vs FADA [44]	3.17×10^{-3}	6.12×10^{-3}
Ours vs CSCL [10]	1.81×10^{-3}	8.75×10^{-4}
Ours vs SS-UDA [33]	3.73×10^{-3}	5.96×10^{-3}
Ours vs PIT [32]	2.45×10^{-3}	8.14×10^{-3}
Ours vs LTIR [23]	6.58×10^{-3}	8.14×10^{-3}
Ours vs ELS-DA [12]	3.62×10^{-3}	4.75×10^{-3}

4.3.3 Effect Investigation of $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$

As depicted in Fig. 6, we investigate the effectiveness of complementary modules $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$ by conducting alternative optimization experiments. Even though equipped with different baseline architectures (*i.e.*, VGG or ResNet), $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$ could reinforce each other mutually and mitigate the domain shift iteratively by constructing a virtuous circle of performance promotion. When $\mathcal{T}_A(\cdot)$ captures where to translate transferable visual appearances, $\mathcal{T}_R(\cdot)$ could further explore how to augment transferable representations across domains and improve the translation performance of $\mathcal{T}_A(\cdot)$ in return. Moreover, Fig. 6 demonstrates the efficient convergence of our KATPAN along the alternative iteration process. Obviously, the performance of our KATPAN converges to a stable value for medical endoscopic dataset [9] when the alternative iteration number is 3.

4.3.4 Parameter Sensitivity

As presented in Fig. 7, we empirically introduce the hyper-parameter discussion to investigate the performance sensitivity of our proposed KATPAN with ResNet as backbone. From the presented results in Fig. 7, we can observe that our proposed model achieves stable adaptation performance even though the hyper-parameters $\{\alpha, \lambda_s(\lambda_t)\}$ and $\{\xi_1, \xi_2\}$ are set in a wide selection range. Furthermore, the best performance on medical endoscopic dataset can be attained when $\alpha = 10, \lambda_s(\lambda_t) = 10^{-4}, \xi_1 = 10^{-3}$ and $\xi_2 = 10^{-2}$. The hyper-parameter experiments of $\{\alpha, \lambda_s(\lambda_t)\}$ in Fig. 7 demonstrate that transferable representations across domains could be purified via the transferability-aware information bottleneck. In addition, the selection experiments of parameters $\{\xi_1, \xi_2\}$ justify that two complementary modules $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$ construct a virtuous circle of performance promotion to capture transferable semantic knowledge with high transfer score.

4.3.5 Investigation of Significant Improvement

To validate whether the performance improvement of our KATPAN is significant when compared with other competing comparison methods, we conduct the t-test evaluation in terms of mIoU via 5 random experiments and present the p value of each comparison in Table 3. The lower p value is, the more confident to determine the significant improvement of each comparison. The p value of the comparison between our proposed KATPAN and ELS-DA [12] is significantly lower than 0.05, which verifies the significant performance improvement of our KATPAN compared with the conference version [12]. More importantly, the presented comparisons in Table 3 (*i.e.*, $p < 0.05$) further illustrate the superiority and effectiveness of our proposed KATPAN.

4.4 Experiments on GTA → Cityscapes Task

4.4.1 Performance Comparison

Table 4 reports the performance comparison between our proposed model and other competing state-of-the-art methods on GTA [35] → Cityscapes [8] task. We have the following conclusions summarized from Table 4: 1) Our proposed model significantly outperforms the previous conference version [12] by an improvement of $2.5\% \sim 2.6\%$ in terms of mIoU, since the information propagation among semantically adjacent representations (*e.g.*, the rider, bike and road have more adjacent semantic relationship.) could be highlighted via our knowledge aggregation-induced transferability perception module to improve performance. 2) When compared with the conference version [12], our self-adaptive pseudo label selection strategy in Eq. (17) pays more attention on hard-to-adapt classes (*e.g.*, train, truck

TABLE 4
Performance comparison between our proposed KATPAN and other competing methods on GTA [35] → Cityscapes [8] task.

Method	Net	road	sidewalk	building	wall	fence	pole	light	sign	veg	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
Source only [40]	VGG	18.1	6.8	64.1	7.3	8.7	21.0	14.9	16.8	45.9	2.4	64.4	41.6	17.5	55.3	8.4	5.0	6.9	4.3	13.8	22.3
Wild [21]		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
CBST [54]		66.7	26.8	73.7	14.8	9.5	28.3	25.9	10.1	75.5	15.7	51.6	47.2	6.2	71.9	3.7	2.2	5.4	18.9	32.4	30.9
SWD [24]		91.0	35.7	78.0	21.6	21.7	**31.8**	**30.2**	25.2	80.2	23.9	74.1	53.1	15.8	79.3	22.1	26.5	1.5	17.2	30.4	39.9
ADV [43]		86.9	28.7	78.7	28.5	25.2	17.1	20.3	10.9	80.0	26.4	70.2	47.1	8.4	81.5	26.0	17.2	18.9	11.7	1.6	36.1
SSF [13]		88.7	32.1	79.5	29.9	22.0	23.8	21.7	10.7	80.8	29.8	72.5	49.5	16.1	82.1	23.2	18.1	3.5	24.4	8.1	37.7
PyCDA [27]		86.7	24.8	80.9	21.4	27.3	30.2	26.6	21.1	**86.6**	28.9	58.8	53.2	17.9	80.4	18.8	22.4	4.1	9.7	6.2	37.2
CrCDA [22]		86.8	37.5	80.4	30.7	18.1	26.8	25.3	15.1	81.5	30.9	72.1	52.8	19.0	82.1	25.4	29.2	10.1	15.8	3.7	39.1
CSCL [10]		89.8	46.1	75.2	30.1	27.9	15.0	20.4	18.9	82.6	39.1	77.6	47.8	17.4	76.2	28.5	33.4	0.5	29.4	30.8	41.4
LSE [41]		86.0	26.0	76.7	33.1	13.2	21.8	30.1	16.5	78.8	25.8	74.7	50.6	18.7	81.8	22.5	30.5	12.3	16.9	25.4	39.0
FADA [44]		92.3	51.1	83.7	33.1	**29.1**	28.5	28.0	21.0	82.6	32.6	**85.3**	55.2	**28.8**	**83.5**	24.4	37.4	0.0	21.1	15.2	43.8
TPLD [39]		83.5	49.9	72.3	17.6	10.7	29.6	28.3	9.0	78.2	20.1	25.7	47.4	13.3	79.6	3.3	19.3	1.3	14.3	**33.5**	34.1
DTST [45]		88.1	35.8	83.1	25.8	23.9	29.2	28.8	**28.6**	83.0	36.7	82.3	53.7	22.8	82.3	26.4	**38.6**	0.0	19.6	17.1	42.4
PIT [32]		86.2	35.0	82.1	31.1	22.1	23.2	29.4	28.5	79.3	31.8	81.9	52.1	23.2	80.4	29.5	26.9	**30.7**	20.5	1.2	41.8
LTIR [23]		**92.5**	**54.5**	**83.9**	**34.5**	25.5	31.0	30.4	18.0	84.1	39.6	83.9	53.6	19.3	81.7	21.1	13.6	17.7	12.3	6.5	42.3
ELS-DA [12]		88.4	47.5	76.8	32.0	28.4	15.6	24.5	22.1	80.8	**40.3**	79.4	46.1	20.4	77.2	25.9	31.8	0.7	26.1	28.3	41.7
Ours-w/oKT		89.5	41.7	78.4	31.6	25.2	18.5	20.1	17.4	76.3	34.8	74.6	41.3	19.8	76.4	**29.8**	26.1	19.2	24.3	27.5	40.7
Ours-w/oTB		88.7	42.0	81.5	30.6	25.9	23.2	26.3	22.7	81.2	36.4	70.1	38.6	21.9	77.8	26.5	31.9	18.3	25.4	27.2	41.9
Ours-w/oPL		90.1	40.3	79.8	32.9	23.4	21.6	24.7	19.1	73.8	32.6	71.3	40.9	22.4	79.3	27.2	28.4	20.5	23.6	29.8	41.1
Ours-w/oAA		89.8	37.6	77.2	30.4	21.9	29.5	30.1	23.7	79.4	31.8	77.6	48.3	19.3	75.8	25.9	33.4	23.1	24.6	30.2	42.6
Ours		90.4	35.7	81.5	32.6	24.9	27.4	29.6	26.8	83.7	34.1	81.5	**55.4**	19.3	78.6	23.8	29.1	25.4	**27.6**	32.5	**44.2**
Source only [19]	ResNet	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
LtA [42]		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
ADV [43]		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
SWLS [9]		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
SSF [13]		90.3	38.9	81.7	24.8	22.9	30.5	37.0	21.2	84.8	38.8	76.9	58.8	30.7	85.7	30.6	38.1	5.9	28.3	36.9	45.4
PyCDA [27]		90.5	36.3	84.4	32.4	28.7	34.6	36.4	31.5	**86.8**	37.9	78.5	62.3	21.5	85.6	27.9	34.8	18.0	22.9	49.3	47.4
CrCDA [22]		92.4	**55.3**	82.3	31.2	29.1	32.5	33.2	35.6	83.5	34.8	84.2	58.9	32.2	84.7	40.6	46.1	2.1	31.1	32.7	48.6
CSCL [10]		89.6	50.4	83.0	35.6	26.9	31.1	37.3	35.1	83.5	40.6	84.0	60.6	**34.3**	80.9	35.1	47.3	0.5	34.5	33.7	48.6
LSE [41]		90.2	40.0	83.5	31.9	26.4	32.6	38.7	37.5	81.0	34.2	84.6	61.6	33.4	82.5	32.8	45.9	6.7	29.1	30.6	47.5
FADA [44]		92.5	47.5	85.1	37.6	**32.8**	33.4	33.8	18.4	85.3	37.7	83.5	63.2	39.7	**87.5**	32.9	47.8	1.6	34.9	39.5	49.2
TPLD [39]		83.2	46.3	74.9	29.8	21.3	33.1	36.0	24.2	86.7	43.2	**87.1**	58.7	24.0	84.0	36.9	49.7	0.0	29.7	0.0	44.7
SS-UDA [33]		90.6	37.1	82.6	30.1	19.1	29.5	32.4	20.6	85.7	40.5	79.7	58.7	31.1	86.3	31.5	48.3	0.0	30.2	35.8	46.3
DTST [45]		90.6	44.7	84.8	34.3	28.7	31.6	35.0	37.6	84.7	**43.3**	85.3	57.0	31.5	83.8	42.6	48.5	1.9	30.4	39.0	49.2
LTIR [23]		**92.9**	55.0	**85.3**	34.2	31.1	34.9	40.7	34.0	85.2	40.1	**87.1**	61.0	31.1	82.5	32.3	42.9	0.3	**36.4**	46.1	50.2
PIT [32]		87.5	43.4	78.8	31.2	30.2	**36.3**	39.9	42.0	79.2	37.1	79.3	**65.4**	37.5	83.2	**46.0**	45.6	25.7	23.5	**49.9**	50.6
ELS-DA [12]		89.4	50.1	83.9	35.9	27.0	32.4	38.6	37.5	84.5	39.6	85.7	61.6	33.7	82.2	36.0	50.4	0.3	33.6	32.1	49.2
Ours-w/oKT		91.3	45.6	81.7	34.3	23.8	33.4	38.2	36.5	83.1	38.5	79.5	58.9	27.2	80.7	39.8	48.2	15.9	25.7	32.3	48.1
Ours-w/oTB		92.1	47.3	84.1	31.9	28.3	29.8	42.3	37.4	81.1	42.3	82.5	60.3	28.9	83.8	36.2	50.1	23.4	27.6	31.3	49.5
Ours-w/oPL		89.4	46.5	82.9	33.7	25.4	29.8	39.1	39.2	82.5	40.4	81.6	55.8	29.4	82.7	33.4	47.5	20.8	29.1	36.4	48.7
Ours-w/oAA		90.2	48.6	83.7	30.4	30.1	27.3	41.8	40.7	83.9	39.2	83.5	58.6	28.2	82.9	41.8	48.1	29.0	31.7	38.6	50.4
Ours		90.8	49.8	85.1	**39.5**	28.4	30.5	**43.1**	34.7	84.9	38.9	84.7	62.6	31.6	85.1	38.7	**51.8**	26.2	35.4	42.6	**51.8**

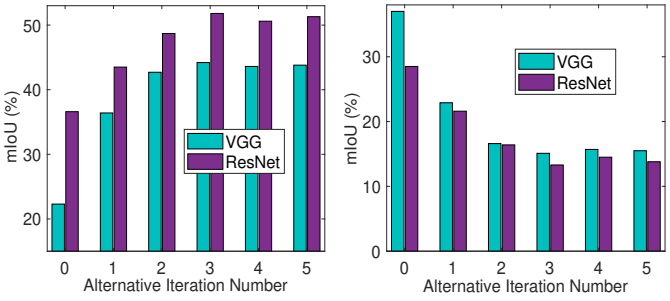


Fig. 10. Effect Investigation of $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$ in terms of mIoU (left) and domain gap (right) on GTA [35] → Cityscapes [8] task.

TABLE 5
Investigation of p value of t-test on GTA [35] → Cityscapes [8] task.

Variants	VGG	ResNet
Ours vs FADA [44]	5.27×10^{-3}	4.83×10^{-3}
Ours vs CSCL [10]	6.19×10^{-3}	3.71×10^{-3}
Ours vs SS-UDA [33]	4.28×10^{-4}	9.51×10^{-4}
DTST [45]	1.94×10^{-2}	4.84×10^{-3}
Ours vs PIT [32]	5.83×10^{-3}	2.87×10^{-4}
Ours vs LTIR [23]	3.72×10^{-4}	4.61×10^{-3}
Ours vs ELS-DA [12]	1.45×10^{-3}	6.24×10^{-4}

and person) and prevents the negative influence of large quantities of pixel categories (*e.g.*, road, sky and building) to generate confident pseudo labels for target examples. 3) Our model outperforms existing advanced frameworks [23], [32], [33], [39], [45] by a significant increase of $1.2\% \sim 17.1\%$ in terms of mIoU, since two complementary modules $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$ with transferability perception capture transferable visual characterizations and transferable semantic representations across domains, especially for those categories with various appearances and diverse modalities (*e.g.*, car, person, bus and rider). The superior adaptation performance of our model is effectively illustrated via some qualitative examples in Fig. 8.

4.4.2 Ablation Study

In this subsection, we introduce the ablation variant experiments on GTA [35] → Cityscapes [8] task to justify the effectiveness of our proposed KATPAN, as shown in Table 4. To be specific, the performance of our model without the knowledge aggregation-induced transferability perception significantly degrades $3.5\% \sim 3.7\%$ in terms of mIoU, which validates the effectiveness of our model to achieve transferability propagation between semantically adjacent representations. Training our model without transferability-aware information bottleneck decreases the performance about 2.3% mIoU. When removing it, we cannot highlight transferable visual characterization across domains via our KATPAN. In addition, the importance of residual attention on attention ($\mathrm{RA}^2\mathrm{B}$) to capture the semantic transferability of transferable representations is effectively illustrated via the performance degradation of $1.4\% \sim 1.6\%$ mIoU. Moreover, after abandoning the pseudo labels of target samples, the decrease of adaptation performance is about 3.1% mIoU. It is indispensable to mitigate the category-wise distribution discrepancy across domains. The confident pseudo labels of target samples are generated along the alternative iteration process of $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$, as depicted in Fig. 9.

4.4.3 Effect Investigation of $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$

This subsection verifies the effectiveness of complementary modules $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$ via extensive evaluation experiments on GTA [35] → Cityscapes [8] task. As depicted in Fig. 10, our model achieves stable convergence performance along with the alternative iteration number of $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$ by exploring where to translate transferable visual appearances and how to augment transferable representations

TABLE 6
Performance comparison between our proposed KATPAN and other competing methods on SYNTHIA [37] → Cityscapes [8] task.

Method	Net	road	sidewalk	building	wall	fence	pole	light	sign	veg	sky	person	rider	car	bus	mbike	bike	mIoU
Source only [40]	VGG	6.4	17.7	29.7	1.2	0.0	15.1	0.0	7.2	30.3	66.8	51.1	1.5	47.3	3.9	0.1	0.0	17.4
Wild [21]		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
CDA [48]		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
DCAN [46]		79.9	30.4	70.8	1.6	0.6	22.3	6.7	23.0	76.9	73.9	41.9	16.7	61.7	11.5	10.3	38.6	35.4
CBST [54]		69.6	28.7	69.5	12.1	0.1	25.4	11.9	13.6	82.0	81.9	49.1	14.5	66.0	6.6	3.7	32.4	35.4
ADV [43]		67.9	29.4	71.9	6.3	0.3	19.9	0.6	2.6	74.9	74.9	35.4	9.6	67.8	21.4	4.1	15.5	31.4
TGCF [7]		90.1	48.6	80.7	2.2	0.2	27.2	3.2	14.3	82.1	78.4	54.4	16.4	82.5	12.3	1.7	21.8	38.5
PyCDA [27]		80.6	26.6	74.5	2.0	0.1	18.1	13.7	14.2	80.8	71.0	48.0	19.0	72.3	22.5	12.1	18.1	35.9
CrCDA [22]		74.5	30.5	78.6	6.6	0.7	21.2	2.3	8.4	77.4	79.1	45.9	16.5	73.1	24.1	9.6	14.2	35.2
CSCL [10]		70.9	30.5	77.8	9.0	0.6	27.3	8.8	12.9	74.8	81.1	43.0	25.1	73.4	34.5	19.5	38.2	39.2
LSE [41]		83.6	39.6	79.3	3.6	0.9	25.3	14.1	26.1	79.4	76.5	51.0	18.1	75.7	22.5	12.0	32.1	40.0
FADA [44]		80.4	35.9	80.9	2.5	0.3	30.4	7.9	22.3	81.8	83.6	48.9	16.8	77.7	31.1	13.5	17.9	39.5
TPLD [39]		81.3	34.5	73.3	11.9	0.0	26.9	0.2	6.3	79.9	71.2	55.1	14.2	73.6	5.7	0.5	41.7	36.0
PIT [32]		81.7	26.9	78.4	6.3	0.2	19.8	13.4	17.4	76.7	74.1	47.5	22.4	76.0	21.7	19.6	27.7	38.1
ELS-DA [12]		74.2	28.9	75.4	10.2	0.5	24.5	6.9	11.8	76.2	80.7	46.1	24.8	74.4	30.3	16.4	40.6	38.9
Ours-w/oKT		84.3	31.7	74.5	3.8	0.4	24.8	10.7	11.4	78.5	77.3	46.3	23.7	74.1	26.4	11.9	24.3	37.8
Ours-w/oTB		88.7	27.3	79.5	5.1	0.7	25.8	12.3	20.1	79.4	75.0	48.5	19.4	71.9	29.3	14.0	22.4	38.7
Ours-w/oPL		86.1	29.7	76.4	6.8	0.6	26.3	14.1	18.2	76.5	73.6	49.3	16.5	73.7	26.6	10.1	24.8	38.1
Ours-w/oAA		89.3	36.5	74.7	7.5	0.5	24.7	9.8	14.6	78.4	71.1	52.3	22.8	76.4	25.6	17.1	33.8	39.7
Ours		86.7	39.6	77.5	6.4	0.7	23.3	11.6	15.2	80.1	78.9	54.8	25.7	79.3	35.1	15.3	30.9	41.3
Source only [19]	ResNet	55.6	23.8	74.6	9.2	0.2	24.4	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	33.5
DCAN [46]		81.5	33.4	72.4	7.9	0.2	20.0	8.6	10.5	71.0	68.7	51.5	18.7	75.3	22.7	12.8	28.1	36.5
CBST [54]		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
ADV [43]		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
SWLS [9]		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
MSL [4]		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
PyCDA [27]		75.5	30.9	83.3	20.8	0.7	32.7	27.3	33.5	84.7	85.0	64.1	25.4	85.0	45.2	21.2	32.0	46.7
CrCDA [22]		86.2	44.9	79.5	8.3	0.7	27.8	9.4	11.8	78.6	86.5	57.2	26.1	76.8	39.9	21.5	32.1	42.9
CSCL [10]		80.2	41.1	78.9	23.6	0.6	31.0	27.1	29.5	82.5	83.2	62.1	26.8	81.5	37.2	27.3	42.9	47.2
LSE [41]		82.9	43.1	78.1	9.3	0.6	28.2	9.1	14.4	77.0	83.5	58.1	25.9	71.9	38.0	29.4	31.2	42.6
FADA [44]		84.5	40.1	83.1	4.8	0.0	34.3	20.1	27.2	84.8	84.0	53.5	22.6	85.4	43.7	26.8	27.8	45.2
TPLD [39]		80.9	44.3	82.2	19.9	0.3	40.6	20.5	30.1	77.2	80.9	60.6	25.5	84.8	41.1	24.7	43.7	47.3
SS-UDA [33]		84.3	37.7	79.5	5.3	0.4	24.9	9.2	8.4	80.0	84.1	57.2	23.0	78.0	38.1	20.3	36.5	41.7
PIT [32]		83.1	27.6	81.5	8.9	0.3	21.8	26.4	33.8	76.4	78.8	64.2	27.6	79.6	31.2	31.0	31.3	44.0
ELS-DA [12]		81.7	43.8	80.1	22.3	0.5	29.4	28.6	21.2	83.4	82.3	63.1	26.2	83.7	34.9	26.3	48.4	47.2
Ours-w/oKT		81.6	37.4	79.8	21.8	0.5	32.2	26.9	27.3	81.7	82.5	60.6	22.4	84.7	32.4	24.2	47.8	46.5
Ours-w/oTB		82.1	40.3	81.9	21.2	0.7	29.4	27.2	30.6	79.5	81.8	66.2	29.4	82.4	34.6	29.1	48.5	47.8
Ours-w/oPL		82.5	39.7	80.5	19.6	0.5	30.7	28.3	29.4	79.5	80.2	61.8	27.1	83.6	30.7	32.9	46.3	47.1
Ours-w/oAA		80.7	43.6	81.8	20.4	0.8	36.2	27.4	30.5	80.2	82.6	64.5	30.7	83.8	38.7	30.1	50.3	48.9
Ours		82.3	40.8	83.7	19.2	1.8	34.6	29.5	32.7	82.9	83.4	67.3	32.8	86.1	41.2	33.5	52.1	50.2

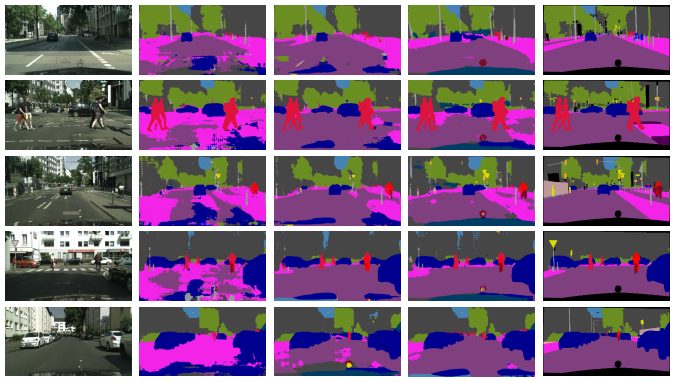


Fig. 11. Qualitative results on SYNTHIA [37] → Cityscapes [8] task.

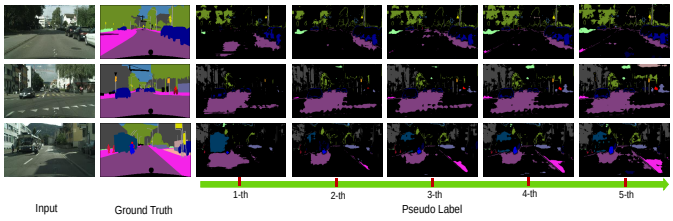


Fig. 12. Pseudo labels on SYNTHIA [37] → Cityscapes [8] task.

across domains. $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$ construct a virtuous circle of performance promotion to minimize the domain discrepancy alternatively by reinforcing each other mutually. The results in Fig. 10 strongly support the stable convergence of our model to minimize domain gap. When the alternative iteration number is 3, our model attains a stable adaptation performance on GTA [35] → Cityscapes [8] task.

4.4.4 *Investigation of Significant Improvement*

In this subsection, we present the t-test validation experiments via 5 random runs to evaluate whether the performance improvement between our model and other competing comparison methods is significant, as introduced in Table 5. When compared with the previous conference version [12], the performance of our proposed model achieves significant improvement, since the p value of Ours vs ELS-DA [12] is significantly lower than 0.05. More importantly, the t-test investigations about significant improvement between our proposed KATPAN and other comparison methods also illustrate the state-of-the-art performance of our model.

4.5 Experiments on SYNTHIA → Cityscapes Task

4.5.1 *Performance Comparison*

This subsection introduces the empirical comparison discussions among our proposed KATPAN and other competing approaches when performing the SYNTHIA [37] → Cityscapes [8] adaptation task, as presented in Table 6. Some essential observations are concluded from Table 6: 1) Our proposed model achieves the state-of-the-art adaptation performance on SYNTHIA [37] → Cityscapes [8] adaptation task, and significantly outperforms the conference version [12] by a large margin of $2.4\% \sim 3.0\%$ in terms of mIoU. 2) Compared with ELS-DA [12], our KATPAN performs better for those hard-to-adapt categories (*e.g.*, person, rider, bus and motorbike), especially for those classes with various modalities and diverse appearances (*e.g.*, bus and car). 3) Two complementary modules $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$ with transferability perception encourage our model to significantly outperform other state-of-the-art models about $1.3\% \sim 21.1\%$ mIoU. It illustrates that transferable visual characterizations and transferable semantic representations across domains could be effectively explored via our KATPAN, even though some hard-to-adapt objects have various appearances among different domains. Some qualitative results on SYNTHIA [37] → Cityscapes [8] task are shown in Fig. 11 to visualize superior adaptation performance of our KATPAN.

TABLE 7
Performance comparison between our proposed KATPAN with ResNet-101 and other competing methods on Cityscapes [8] → NTHU [5] task.

City	Method	road	sidewalk	building	light	sign	veg	sky	person	rider	car	bus	mbike	bike	mIoU
Rome	Source only [19]	83.9	34.3	87.7	13.0	41.9	84.6	92.5	37.7	22.4	80.8	38.1	39.1	5.3	50.9
	NMD [5]	79.5	29.3	84.5	0.0	22.2	80.6	82.8	29.5	13.0	71.7	37.5	25.9	1.0	42.9
	CBST [54]	**87.1**	**43.9**	89.7	14.8	**47.7**	85.4	90.3	45.4	26.6	85.4	20.5	49.8	10.3	53.6
	LtA [42]	83.9	34.2	88.3	18.8	40.2	86.2	93.1	47.8	21.7	80.9	47.8	48.3	8.6	53.8
	MSL [4]	82.9	32.6	86.7	20.7	41.6	85.0	93.0	47.2	22.5	82.2	**53.8**	50.5	9.9	54.5
	SSF [13]	84.2	38.4	87.4	**23.4**	43.0	85.6	88.2	50.2	23.7	80.6	38.1	51.6	8.6	54.1
	CSCL [10]	85.7	36.5	**92.1**	19.4	42.6	84.8	**95.0**	46.9	**28.3**	79.4	40.5	**54.2**	7.5	54.8
	FADA [44]	84.9	35.8	88.3	20.5	40.1	85.9	92.8	**56.2**	23.2	83.6	31.8	53.2	14.6	54.7
	Ours	86.1	39.4	90.6	22.5	43.1	**86.3**	93.4	53.5	24.3	**85.9**	32.3	50.8	**14.7**	**55.6**
Rio	Source only [19]	76.6	47.3	82.5	12.6	22.5	77.9	86.5	43.0	19.8	74.5	36.8	29.4	16.7	48.2
	NMD [5]	74.2	43.9	79.0	2.4	7.5	77.8	69.5	39.3	10.3	67.9	41.2	27.9	10.9	42.5
	CBST [54]	**84.3**	55.2	**85.4**	**19.6**	**30.1**	80.5	77.9	55.2	28.6	**79.7**	33.2	37.6	11.5	52.2
	LtA [42]	76.2	44.7	84.6	9.3	25.5	81.8	87.3	55.3	32.7	74.3	28.9	43.0	27.6	51.6
	MSL [4]	76.9	48.8	85.2	13.8	18.9	81.7	**88.1**	54.9	34.0	76.8	39.8	44.1	29.7	53.3
	SSF [13]	74.2	43.7	82.5	10.3	21.7	79.4	86.7	55.9	36.1	74.9	33.7	**52.6**	33.7	52.7
	CSCL [10]	79.5	52.7	83.6	12.4	23.0	80.9	79.7	56.1	**37.7**	72.4	36.0	51.6	34.1	53.8
	FADA [44]	80.6	53.4	84.2	5.8	23.0	78.4	87.7	60.2	26.4	77.1	37.6	53.7	**42.3**	54.7
	Ours	81.4	**55.8**	84.7	11.3	26.3	**82.4**	86.1	**62.3**	27.5	75.2	**41.8**	50.7	39.3	**55.8**
Tokyo	Source only [19]	82.9	31.3	78.7	14.2	24.5	81.6	89.2	48.6	33.3	70.5	7.7	11.5	45.9	47.7
	NMD [5]	83.4	35.4	72.8	12.3	12.7	77.4	64.3	42.7	21.5	64.1	**20.8**	8.9	40.3	42.8
	CBST [54]	85.2	33.6	80.4	8.3	**31.1**	83.9	78.2	53.2	28.9	72.7	4.4	27.0	47.0	48.8
	LtA [42]	81.5	26.0	77.8	17.8	26.8	82.7	90.9	55.8	38.0	72.1	4.2	24.5	50.8	49.9
	MSL [4]	81.2	30.1	77.0	12.3	27.3	82.8	89.5	58.2	32.7	71.5	5.5	**37.4**	48.9	50.5
	SSF [13]	82.1	27.4	78.0	**18.4**	26.6	83.0	90.8	57.1	35.8	72.0	4.6	27.3	**52.8**	50.4
	CSCL [10]	83.1	35.5	81.2	15.8	24.9	81.3	86.4	58.8	**39.2**	68.1	6.7	30.4	51.2	51.0
	FADA [44]	**85.8**	39.5	76.0	14.7	24.9	**84.6**	91.7	**62.2**	27.7	71.4	3.0	29.3	56.3	51.3
	Ours	84.8	**41.3**	**81.7**	17.2	26.7	78.6	**91.8**	58.4	29.7	**74.2**	8.1	31.7	53.1	**52.1**
Taipei	Source only [19]	83.5	33.4	86.6	12.7	16.4	77.0	92.1	17.6	13.7	70.7	37.7	44.4	18.5	46.5
	NMD [5]	78.6	28.6	80.0	13.1	7.6	68.2	82.1	16.8	9.4	60.4	34.0	26.5	9.9	39.6
	CBST [54]	**86.1**	35.2	84.2	15.0	**22.2**	75.6	74.9	22.7	**33.1**	**78.0**	37.6	**58.0**	30.9	50.3
	LtA [42]	81.7	29.5	85.2	26.4	15.6	76.7	91.7	31.0	12.5	71.5	41.1	47.3	27.7	49.1
	MSL [4]	80.7	32.5	85.5	**32.7**	15.1	78.1	91.3	32.9	7.6	69.5	**44.8**	52.4	34.9	50.6
	SSF [13]	84.5	35.3	86.4	17.7	16.9	77.7	91.3	31.8	22.3	73.7	41.1	55.9	28.5	51.0
	CSCL [10]	83.4	33.7	**87.5**	24.3	17.2	75.8	90.6	33.2	24.1	75.3	35.8	56.4	31.2	51.4
	FADA [44]	86.0	42.3	86.1	6.2	20.5	78.3	**92.7**	**47.2**	17.7	72.2	37.2	54.3	44.0	52.7
	Ours	82.8	**42.7**	85.3	23.8	21.3	**79.6**	91.2	42.8	21.6	67.1	40.5	50.7	**45.4**	**53.4**

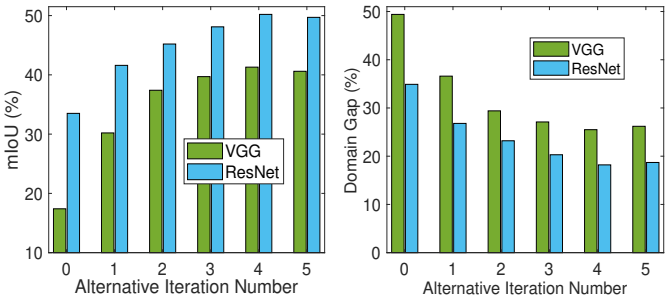


Fig. 13. Effect Investigation of $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$ in terms of mIoU (left) and domain gap (right) on SYNTHIA [37] → Cityscapes [8] task.

4.5.2 Ablation Study

We report the ablation experiments of our model performing on SYNTHIA [37] → Cityscapes [8] adaptation task to verify the effectiveness of each component, as shown in Table 6. Generally, the performance of our KATPAN undergoes a significant decrease of $1.3\% \sim 3.7\%$ in terms of mIoU when we remove any component of our model. It demonstrates the necessity and importance of each module to attain the best adaptation performance cooperatively. When the knowledge aggregation-induced transferability perception module is abandoned, the performance of our model has the largest degradation (*i.e.*, $3.5\% \sim 3.7\%$ mIoU), which justifies the effectiveness of this module to achieve transferability information propagation between semantically adjacent representations. Furthermore, compared with the conference version [12], the self-adaptive pseudo label selection strategy is indispensable to minimize the distribution discrepancy across domains, where the generation process on SYNTHIA [37] → Cityscapes [8] task is introduced in Fig. 12.

4.5.3 Effect Investigation of $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$

As depicted in Fig. 13, we intend to investigate the cooperation effect of two complementary modules $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$ on SYNTHIA [37] → Cityscapes [8] task. From the results in Fig. 13, we observe that $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$ could promote each other mutually to achieve stable convergence performance along the alternative iteration process. This observation effectively supports the convergence analysis that two complementary modules $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$ could jointly encourage our model to achieve a small target error.

4.6 Experiments on Cityscapes → NTHU Task

In this subsection, we report the performance comparison between our KATPAN with ResNet-101 and other competing methods on Cityscapes [8] → NTHU [5] task in Table 7, and introduce the ablation studies of each component of our model in Fig. 14. From the presented results, we can conclude that: 1) Our proposed model achieves the best adaptation performance in all cities, which significantly outperforms existing comparison methods about $0.7 \sim 13.3\%$ in terms of mIoU. It validates the effectiveness of our KATPAN to explore transferable visual appearances and transferable representations across domains. 2) Removing any one of modules in our proposed model results in a large performance decrease of $0.7\% \sim 3.5\%$ mIoU, as shown in Fig. 14. It illustrates that all components of our model play an indispensable role in purifying transferable knowledge while neglecting untransferable representations. 3) Our KATPAN achieves remarkable adaptation performance for those hard-to-adapt categories with different appearances and modalities across domains (*e.g.*, bus, car and person), since two complementary modules $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$ could alternatively capture transferable knowledge and abandon the negative transfer of untransferable representations.

4.7 Qualitative Analysis of Transferability Perception

As shown in Table 8, we conduct more comparison experiments to validate the effectiveness of our proposed

TABLE 8
Performance investigation (mIoU) about transferability perception, adaptive update of $\{\lambda_s, \lambda_t\}$ via using $\mathcal{L}_b^s$ and $\mathcal{L}_b^t$, Gaussian noise in $\mathcal{T}_A(\cdot)$, category-wise prototypes and pseudo label selection strategy.

Tasks	Net	Ours-wTQ	△(%)	Ours-w/oAU	△(%)	Ours-wUN	Ours-wRN	Ours-wGN	Ours-w/oGP	△(%)	Ours-w/oLR	Ours-w/oER	Ours
Medical Endoscopic Dataset [9]	VGG	54.6	↓5.2	58.8	↓1.0	57.3	57.8	58.2	58.9	↓0.9	58.8	59.2	**59.8**
GTA [35] → Cityscapes [8]		40.1	↓4.1	43.4	↓0.8	41.7	42.5	42.2	43.2	↓1.0	43.8	43.6	**44.2**
SYNTHIA [37] → Cityscapes [8]		36.8	↓4.5	40.6	↓0.7	39.0	39.7	39.2	40.5	↓0.8	40.3	40.6	**41.3**
Medical Endoscopic Dataset [9]	ResNet	62.5	↓4.2	65.6	↓1.1	64.2	64.8	65.3	65.9	↓0.8	66.1	65.7	**66.7**
GTA [35] → Cityscapes [8]		57.3	↓4.5	50.9	↓0.9	48.9	49.7	49.3	50.6	↓1.2	51.2	51.4	**51.8**
SYNTHIA [37] → Cityscapes [8]		46.9	↓3.3	49.5	↓0.7	48.3	49.4	48.7	49.3	↓0.9	49.8	49.3	**50.2**
Cityscapes [8] → NTHU (Rome) [5]		52.1	↓3.5	54.7	↓0.9	54.1	54.7	54.8	55.0	↓0.6	55.1	55.3	**55.6**
Cityscapes [8] → NTHU (Rio) [5]		51.9	↓3.9	55.1	↓0.7	53.5	54.1	54.3	55.1	↓0.7	55.4	55.2	**55.8**
Cityscapes [8] → NTHU (Tokyo) [5]		48.7	↓3.4	51.3	↓0.8	50.2	50.5	50.8	51.5	↓0.6	51.4	51.6	**52.1**
Cityscapes [8] → NTHU (Taipei) [5]		50.1	↓3.3	52.7	↓0.7	51.6	51.9	52.2	52.6	↓0.8	52.8	53.1	**53.4**

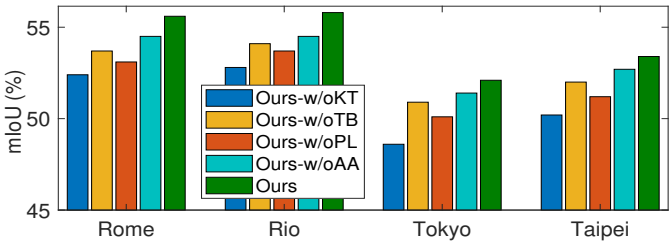


Fig. 14. Ablation investigation on Cityscapes [8] → NTHU [5] task.

knowledge aggregation-induced transferability perception. We modify the transferability quantification strategy in Eqs. (1) and (2) as $Q_{F_s} = |\mathcal{I}(D_F(F_i^s;\theta_{D_F})) - 0.5|$ and $Q_{F_t} = |\mathcal{I}(D_F(F_j^t;\theta_{D_F})) - 0.5|$, and denote this variant as Ours-wTQ. The performance of Ours-wTQ decreases by 3.3%∼5.2% in terms of mIoU, when compared with Ours. It validates the effectiveness and superiority of our proposed knowledge aggregation-induced transferability perception. Furthermore, Ours-wTQ unreasonably assumes that all representations across domains are transferable by setting the threshold as 0.5. Such strategy introduces false quantified transferability perception for transferable, already transferred, or untransferable representations, which heavily degrades the adaptation performance. When compared with it, our transferability perception strategy could quantify accurate transferability for cross-domain representations.

4.8 Qualitative Analysis of Adaptive Update of $\{\lambda_s, \lambda_t\}$

This subsection investigates the effectiveness of using $\mathcal{L}_b^s$ and $\mathcal{L}_b^t$ to adaptively update $\{\lambda_s, \lambda_t\}$, as shown in Table 8. We replace the adaptive update strategy of $\{\lambda_s, \lambda_t\}$ with a fixed initialization value (*i.e.*, without using $\mathcal{L}_b^s$ and $\mathcal{L}_b^t$), and denote this variant as Ours-w/oAU. From the presented results in Table 8, we observe that the performance of Ours-w/oAU degrades $0.7\% \sim 1.1\%$ in terms of mIoU, when compared with Ours. It illustrates the effectiveness to adaptively update $\{\lambda_s, \lambda_t\}$ via the network itself. Moreover, the proposed optimization strategy could perform a specific information bottleneck constraint on latent space to adaptively purify adaptation-dependent representations.

4.9 Qualitative Analysis of Gaussian Noise in $\mathcal{T}_A(\cdot)$

We introduce comparison experiments to validate the effectiveness of Gaussian noisy in $\mathcal{T}_A(\cdot)$, as shown in Table 8. We respectively replace Gaussian noise with uniform noise, rayleigh noise and gamma noise; and denote these variants as Ours-wUN, Ours-wRN and Ours-wGN. From the presented results in Table 8, we observe that the performances of Ours-wUN, Ours-wRN and Ours-wGN degrade 1.2%∼2.9% in terms of mIoU. This observation experimentally illustrates that the latent representations will be more discriminative if they are closer to a Gaussian distribution, as introduced in [1], [34]. The information bottleneck constraint between Gaussian noise $G(z)$ and the semantic representation extracted via $E_A(\cdot)$ in the latent feature space could effectively explore adaptation-dependent transferable knowledge, while neglecting adaptation-independent untransferable knowledge.

4.10 Qualitative Analysis of Category-Wise Prototypes

In this subsection, as shown in Table 8, we present the evaluation experiments about global category-wise prototypes to validate the effectiveness of transferability knowledge graph. Denote our proposed model without using global category-wise prototypes to structure transferability knowledge graph as Ours-w/oGP. When compared with Ours, the performance of Ours-w/oGP decreases by 0.6%∼1.2% in terms of mIoU. The degradation performance validates the effectiveness to structure both global category-wise prototypes and semantic representations as transferability knowledge graph. Moreover, by incorporating with global category-wise prototypes, the transferability knowledge graph could characterize category-aware transferability correlations between semantically adjacent representations and global category-wise prototypes from intra and inter domains. It also promotes the transferability propagation between global category-wise prototypes and semantically adjacent representations.

4.11 Qualitative Analysis of Pseudo Label Selection

Table 8 introduces the qualitative analysis of self-adaptive pseudo label selection strategy on several benchmark datasets when the backbone of our proposed model is VGG [40] or ResNet [19]. In general, Ours-w/oLR and Ours-w/oER are the abbreviations of self-adaptive pseudo label selection strategy without label regularizer and entropy regularizer, respectively. From the presented results in Table 8, we can conclude that both label regularizer and entropy regularizer can cooperate well to mine confident pseudo labels and achieve the better adaptation performance. When compared with Ours, the large performance degradation (mIoU) of Ours-w/oLR validates the effectiveness to smooth the confident prediction of pseudo labels and further restrain the noise generation. Moreover, the self-adaptive pseudo

TABLE 9
Performance (mIoU) of our model with transformer as backbone.

Tasks	VGG	ResNet	Transformer
Medical Endoscopic Dataset [9]	59.8	66.7	**67.0**
GTA [35] → Cityscapes [8]	44.2	51.8	**51.9**
SYNTHIA [37] → Cityscapes [8]	41.3	50.2	**50.4**
Cityscapes [8] → NTHU (Rome) [5]	-	**55.6**	55.4
Cityscapes [8] → NTHU (Rio) [5]	-	55.8	**56.1**
Cityscapes [8] → NTHU (Tokyo) [5]	-	**52.1**	51.8
Cityscapes [8] → NTHU (Taipei) [5]	-	53.4	**53.5**

label selection strategy without using entropy regularizer (*i.e.*, Ours-w/oER) significantly decreases the adaptation performance in terms of mIoU. It illustrates that the entropy regularizer can effectively minimize the uncertainty of unconfident prediction and encourage the confident pseudo label prediction to be sharper.

4.12 Qualitative Analysis of Transformer Backbone

As shown in Table 9, this subsection presents the performance investigation of our model when the backbone architecture (*i.e.*, VGG [40] or ResNet [19]) is replaced with the recent popular Transformer [51]. Specifically, as introduced in [51], we first split the input image into a sequence of size-fixed image patches, then forward the flattened pixel vectors of each image patch into a linear embedding layer, and add position embedding to obtain a sequence of feature embedding vectors. The transformer encoder takes the embedding sequence as input to obtain discriminative feature representation via the self-attention mechanism. Given the learned feature representations from transformer encoder, the multi-level feature aggregation strategy in [51] is employed to recover original image resolution for segmentation task.

From the performance in Table 9, we can conclude that the performance (mIoU) of our proposed model with transformer as backbone outperforms the VGG or ResNet backbone in most cases. It validates that our model with transformer as backbone could effectively explore shared long-range dependency knowledge across domains in semantic segmentation task. The transformer backbone models global interactions of embedding sequences to learn shared feature representations across domains, which effectively mitigates cross-domain distribution discrepancy.

5 Conclusion

In this paper, we propose a novel Knowledge Aggregation-induced Transferability Perception Adaptation Network (KATPAN) to explore where and how to capture transferable visual characterizations and semantic representations for unsupervised domain adaptation. To be specific, the knowledge aggregation-induced transferability perception (KATP) module is designed to quantify the adaptation contributions of semantic representations across domains via transferability information propagation from global category-wise prototypes. Furthermore, two complementary modules (*i.e.*, $\mathcal{T}_A(\cdot)$ and $\mathcal{T}_R(\cdot)$) are developed to alternatively explore where to translate transferable visual appearances and how to augment transferable representations across domains, by constructing a virtuous circle of performance promotion. Comprehensive comparison experiments on several representative datasets strongly demonstrate the state-of-the-art performance of our KATPAN, when compared with other competing approaches. In the future, we will extend our proposed KATPAN into other challenging visual tasks such as domain-adaptive object detection and multi-source lifelong domain adaptation.

References

[1] Alex Alemi, Ian Fischer, Josh Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2017.

[2] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(85):2399–2434, 2006.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[4] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[5] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[6] Haoang Chi, Feng Liu, Wenjing Yang, Long Lan, Tongliang Liu, Bo Han, William K. Cheung, and James T. Kwok. TOHAN: A one-step approach towards few-shot hypothesis adaptation. 2021.

[7] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

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

[9] Jiahua Dong, Yang Cong, Gan Sun, and Dongdong Hou. Semantic-transferable weakly-supervised endoscopic lesions segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10711–10720, October 2019.

[10] Jiahua Dong, Yang Cong, Gan Sun, Yuyang Liu, and Xiaowei Xu. Cscl: Critical semantic-consistent learning for unsupervised domain adaptation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European Conference on Computer Vision – ECCV 2020*, pages 745–762, Cham, 2020. Springer International Publishing.

[11] Jiahua Dong, Yang Cong, Gan Sun, Yunsheng Yang, Xiaowei Xu, and Zhengming Ding. Weakly-supervised cross-domain adaptation for endoscopic lesions segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2020.

[12] Jiahua Dong, Yang Cong, Gan Sun, Bineng Zhong, and Xiaowei Xu. What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4022–4031, June 2020.

[13] Liang Du, Jingang Tan, Hongye Yang, Jianfeng Feng, Xiangyang Xue, Qibao Zheng, Xiaoqing Ye, and Xiaolin Zhang. Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[14] Zhen Fang, Jie Lu, Anjin Liu, Feng Liu, and Guangquan Zhang. Learning bounds for open-set learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 3122–3132. PMLR, 2021.

[15] Zhen Fang, Jie Lu, Feng Liu, Junyu Xuan, and Guangquan Zhang. Open set domain adaptation: Theoretical bound and algorithm. *IEEE Transactions on Neural Networks and Learning Systems*, abs/1907.08375, 2019.

[16] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 06–11 Aug 2017.

[17] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016.

[18] Golnaz Ghiasi and Charless Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *European Conference on Computer Vision*, volume 9907, pages 519–534, October 2016.

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

[20] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1989–1998. PMLR, 10–15 Jul 2018.

[21] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

[22] Jiaxing Huang, Shijian Lu, Dayan Guan, and Xiaobing Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European Conference on Computer Vision – ECCV 2020*, pages 705–722, Cham, 2020. Springer International Publishing.

[23] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[24] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[25] Jingjing Li, Erpeng Chen, Zhengming Ding, Lei Zhu, Ke Lu, and Heng Tao Shen. Maximum density divergence for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

[26] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[27] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[28] Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J. Sutherland. Learning deep kernels for non-parametric two-sample tests. In *ICML*, volume 119, pages 6316–6326. PMLR, 2020.

[29] Sifei Liu, Shalini De Mello, Jinwei Gu, Guangyu Zhong, Ming-Hsuan Yang, and Jan Kautz. Learning affinity via spatial propagation networks. In *Advances in Neural Information Processing Systems 30*, pages 1520–1530. 2017.

[30] Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Semantic image segmentation via deep parsing network. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[31] Mingsheng Long, Jianmin Wang, Guiguang Ding, Sinno Jialin Pan, and Philip S. Yu. Adaptation regularization: A general framework for transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 26(5):1076–1089, 2014.

[32] Fengmao Lv, Tao Liang, Xiang Chen, and Guosheng Lin. Cross-domain semantic segmentation via domain-invariant interactive relation transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[33] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[34] Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel, and Sergey Levine. Variational discriminator bottleneck: Improving imitation learning, inverse RL, and GANs by constraining information flow. In *International Conference on Learning Representations*, 2019.

[35] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS*, pages 102–118. Springer International Publishing, 2016.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint arXiv:1505.04597*, August 2015.

[37] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio 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.

[38] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651, April 2017.

[39] Inkyu Shin, Sanghyun Woo, Fei Pan, and In So Kweon. Two-phase pseudo label densification for self-training based domain adaptation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European Conference on Computer Vision – ECCV 2020*, pages 532–548, Cham, 2020. Springer International Publishing.

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[41] M. Naseer Subhani and Mohsen Ali. Learning from scale-invariant examples for domain adaptation in semantic segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 290–306, Cham, 2020. Springer International Publishing.

[42] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[43] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Perez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[44] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European Conference on Computer Vision – ECCV 2020*, pages 642–659, Cham, 2020. Springer International Publishing.

[45] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S. Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[46] Zuxuan Wu, Xintong Han, Yen-Liang Lin, Mustafa Gokhan Uzunbas, Tom Goldstein, Ser Nam Lim, and Larry S. Davis. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[47] Hanchao Yu, Shanhui Sun, Haichao Yu, Xiao Chen, Honghui Shi, Thomas S. Huang, and Terrence Chen. Foal: Fast online adaptive learning for cardiac motion estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[48] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[49] Yiyang Zhang, Feng Liu, Zhen Fang, Bo Yuan, Guangquan Zhang, and Jie Lu. Clarinet: A one-step approach towards budget-friendly unsupervised domain adaptation. In Christian Bessiere, editor, *IJCAI*, pages 2526–2532. ijcai.org, 2020.

[50] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[51] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang,

Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6881–6890, June 2021.

[52] Li Zhong, Zhen Fang, Feng Liu, Jie Lu, Bo Yuan, and Guangquan Zhang. How does the combined risk affect the performance of unsupervised domain adaptation approaches? In *AAAI*, pages 11079–11087. AAAI Press, 2021.

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

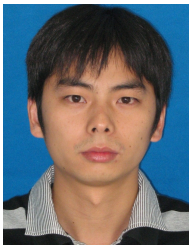
[54] Yang Zou, Zhiding Yu, B.V.K. Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *The European Conference on Computer Vision (ECCV)*, September 2018.

Jiahua Dong is currently a Ph. D candidate in the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences. He received the B.S. degree from Jilin University in 2017. He also has some top-tier journal/conference papers accepted at TPAMI, CVPR, ICCV, ECCV, AAAI, et al. He has served as a PC member for NeurIPS/CVPR/ICCV/ICLR/AAAI/IJCAI, and a reviewer for over 10 international journals such as IEEE TNNLS/TMI/TCYB/JAS/TCSVT. His current research interests include transfer learning, incremental learning, robotic vision 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, computer 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-M'20) is an associate 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, Shenyang Institute of Automation, 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. He also has some top-tier conference papers accepted at CVPR, ICCV, ECCV, AAAI, IJCAI, ICDM et al. His current research interests include lifelong machine learning, multitask learning, medical data analysis, deep learning and 3D computer vision.

Zhen Fang received his M.Sc. degree in pure mathematics from the School of Mathematical Sciences Xiamen University, Xiamen, China, in 2017. He is working toward a PhD degree with the Faculty of Engineering and Information Technology, University of Technology Sydney, Australia. His research interests include transfer learning and domain adaptation. He is a Member of the AAII, University of Technology Sydney.



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 Science, Tulane University since 2021. Prior that, he was a faculty member affiliated with Department of Computer, Information and Technology, Indiana University-Purdue University Indianapolis. 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 (ACM MM 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.