Pattern Recognition 112 (2021) 107764

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Scale variance minimization for unsupervised domain adaptation in image segmentation

Dayan Guan, Jiaxing Huang, Shijian Lu*, Aoran Xiao

Nanyang Technological University, 50 Nanyang Avenue, 639798, Singapore

ARTICLE INFO

Article history:
Received 18 August 2020
Revised 8 October 2020
Accepted 21 November 2020
Available online 5 December 2020

Keywords:
Unsupervised domain adaptation
Image segmentation
Semantic structure
Variance minimization
Adversarial learning

ABSTRACT

We focus on unsupervised domain adaptation (UDA) in image segmentation. Existing works address this challenge largely by aligning inter-domain representations, which may lead over-alignment that impairs the semantic structures of images and further target-domain segmentation performance. We design a scale variance minimization (SVMin) method by enforcing the intra-image semantic structure consistency in the target domain. Specifically, SVMin leverages an intrinsic property that simple scale transformation has little effect on the semantic structures of images. It thus introduces certain supervision in the target domain by imposing a scale-invariance constraint while learning to segment an image and its scale-transformation concurrently. Additionally, SVMin is complementary to most existing UDA techniques and can be easily incorporated with consistent performance boost but little extra parameters. Extensive experiments show that our method achieves superior domain adaptive segmentation performance as compared with the state-of-the-art. Preliminary studies show that SVMin can be easily adapted for UDA-based image classification.

1. Introduction

Image segmentation, which aims to assign class labels to every pixel of an input image, has been a longstanding challenge in computer vision research. Fully supervised approaches [17,36,38] have achieved great successes at the price of large-scale densely-annotated datasets [3,5] that are prohibitively expensive and time-consuming to collect. One way of circumventing this constraint is to leverage synthesized images with self-contained annotations [21,22] for network training. On the other hand, the models trained using synthesized images usually undergo a drastic performance drop while applied to natural scene images [28,30].

Unsupervised domain adaptation (UDA) refers to the task of training a model on labelled data of a source domain for achieving good performance in a target domain, with access to only unlabelled data in the target domain. Based on the theoretical insight that minimizing the inter-domain discrepancy lowers the upper bound of errors in the target domain [1], state-of-the-art methods [32,35,37] address the UDA challenge largely by minimizing the discrepancy between the source and target domains. Although these prior works have achieved quite promising progress, they may impair the semantic structures of images in the target domain while striving to align the inter-domain representations desperately. In another word, the brute-force inter-domain alignment may undesirably damage the integrity of semantic structures of target domain images which is critically important to image segmentation.

We propose a scale variance minimization (SVMin) technique that leverages a scale-invariance constraint for preserving the semantic structures of images in the target domain while performing the inter-domain alignment. The idea is to leverage an intrinsic property of images that scale transformations have minimal effect on the semantic structures of images. The scale-invariance constraint thus provides certain supervision within the target domain, which is enforced through segmenting an image and its scale transformation concurrently as illustrated in Fig. 1. SVMin is orthogonal and complementary to the minimization of inter-domain discrepancy by preventing over-alignment that often disintegrates the semantic structures of images in the target domain. It has a unique feature that it can work with most existing UDA-based image segmentation networks (as a plug-in) with consistent performance improvement but little extra parameters. Additionally, SVMin can work well in UDA-based image classification.

The main contributions of this work are threefold. First, it identifies a scale-invariance constraint, an orthogonal component to the classical inter-domain alignment that leverages certain target-domain consistency for optimal unsupervised domain adaptation. Second, it proposes a scale variance minimization (SVMin) tech-

* Corresponding author.
E-mail addresses: dayan.guan@ntu.edu.sg (D. Guan), jiaxing.huang@ntu.edu.sg (J. Huang), shijian.lu@ntu.edu.sg (S. Lu), aoran.xiao@ntu.edu.sg (A. Xiao).

https://doi.org/10.1016/j.patcog.2020.107764
0031-3203/

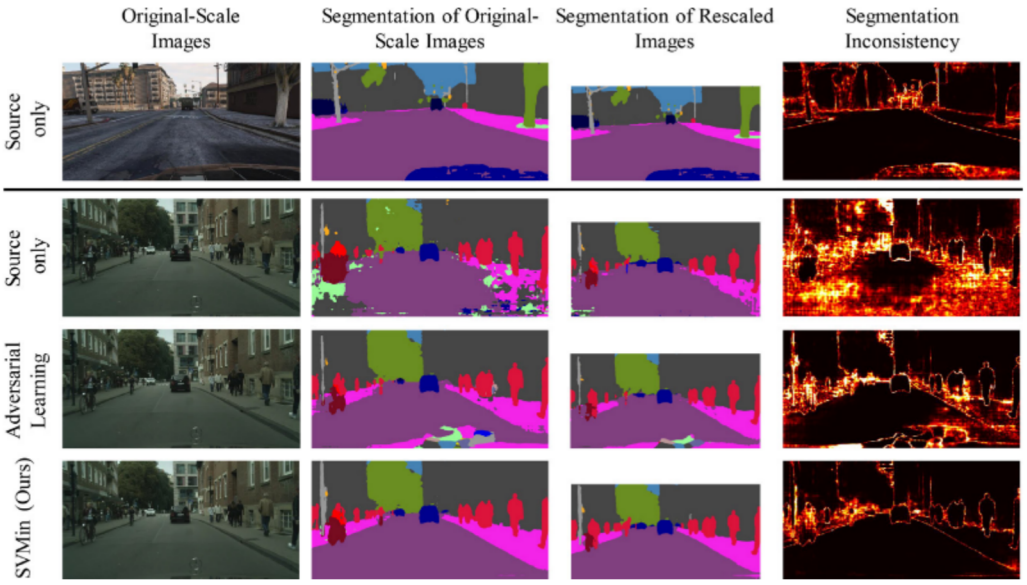

Fig. 1. The proposed SVMin method introduces certain supervision in the unlabelled target domain, i.e. semantic segmentation of the same image at different image scales should be largely the same: The *Source only* trained using the labelled source-domain images demonstrates good consistency across segmentation at the original and rescaled source-domain images (**Row 1**), but it produces poor consistency while applied to target-domain images directly without adaptation (**Row 2**). The prevalent *Adversarial Learning* [30] performs inter-domain alignment with better segmentation and consistency in the target domain, though it often over-aligns the two domains and disintegrates semantic structures of images in the target domain (**Row 3**). SVMin introduces a scale-invariance constraint by enforcing the consistency of semantic structures across image scales, which achieves better results in the target domain (**Row 4**).

nique that achieves superior domain adaptive segmentation performance by enforcing the consistency of semantic image structures in the target domain. Third, it demonstrates that SVMin can work with most existing unsupervised domain adaptation techniques with consistent performance boost, and it is generic and can be easily adapted to other domain adaptive tasks such as image classification.

The remainder of the paper is organized as follows. We review existing works for UDA-base image segmentation and multi-view learning in Section 2. The details of our proposed SVMin methods and the correlate theoretical insights into SVMin are presented in Section 3. An extensive experimental evaluation of our method in UDA-based image segmentation and classification is provided in Sections 4 and 5 concludes this paper.

2. Related works

Unsupervised domain adaptation (UDA) refers to the task of adapting a model from a labelled source domain to an unlabelled target domain optimally [4,14]. One major driving force of UDA research is for mitigating the constraint of data collection and annotation in deep neural network training, mainly suppressing the data discrepancy among different domains. UDA has been widely studied in various computer vision tasks such as image classification [12,20,40], object detection [33], and image segmentation [11,30,35]. Our methods focus on the task of UDA-based image segmentation which will be carefully reviewed in the following part.

UDA-based image segmentation aims to design UDA techniques to handle the domain adaptive image segmentation problem. Most existing methods endeavour to suppress the inter-domain discrepancy and align the feature representation across domains, through either one-stage style transfer [7], adversarial learning [28–30], self-training [39] or a two-stage processing procedure [9,32,37]. Hong et al. [7] designed a conditional generator to translate features of source images to target images and a discriminator to distinguish them. Vu et al. [30] introduced an entropy-based adversarial training approach to achieve intra-domain entropy minimization and inter-domain alignment. Zou et al. [39] proposed a confidence regularized self-training framework formulated as regularized self-training loss minimization. Zhang et al. [37] adapted pretrained model via AdaptSeg [28] to the target domain through category-wise feature alignment guided by category anchors. Kim [9] et al. encoded the texture in source domain with a style transfer algorithm and translated the source images with image translation networks. Wang et al. [32] minimized the distance of the closest stuff and instance features between source and target domain in the intra-domain self-training stage. Though the inter-domain alignment achieves certain success in UDA-based image segmentation, it tends to lack regulations and damage the integrity of semantic structures of target-domain images and further leads to degraded image segmentation.

Multi-view learning refers to the method that learners are trained alternately on two or multiple different views with confident predictions from the unlabeled target data. In the field of UDA, these methods [23–25] are capable to generate pseudo labels for unlabeled target data, which enables direct measurement and minimizing the task loss (e.g., cross entropy loss in classification/detection/segmentation) on unlabelled target domain. Currently, multi-view learning methods enforce multiple classifiers to become diverse/distinct by diversifying the learned parameters (e.g., kernel weights), through adversarial dropout [24], classifier discrepancy maximization [25] or asymmetric classifier tri-training [23], etc. In addition to utilizing multi-view training for unlabeled target data pseudo labels generation, Saito et al. [25] proposed to maximize the consensus of multiple classifiers for UDA. Different from previous multi-view learning methods that create multiple views on feature space by employing multiple classifiers, we directly create multiple views on input space by resizing the input image with different ratios.

Our SVMin introduces a scale-invariance constraint (as supervision in the target domain) to regulate the inter-domain alignment, targeting to maintain the integrity of semantic structures of target-domain images while aligning feature representation between the source and target domains. To the best of our knowl-

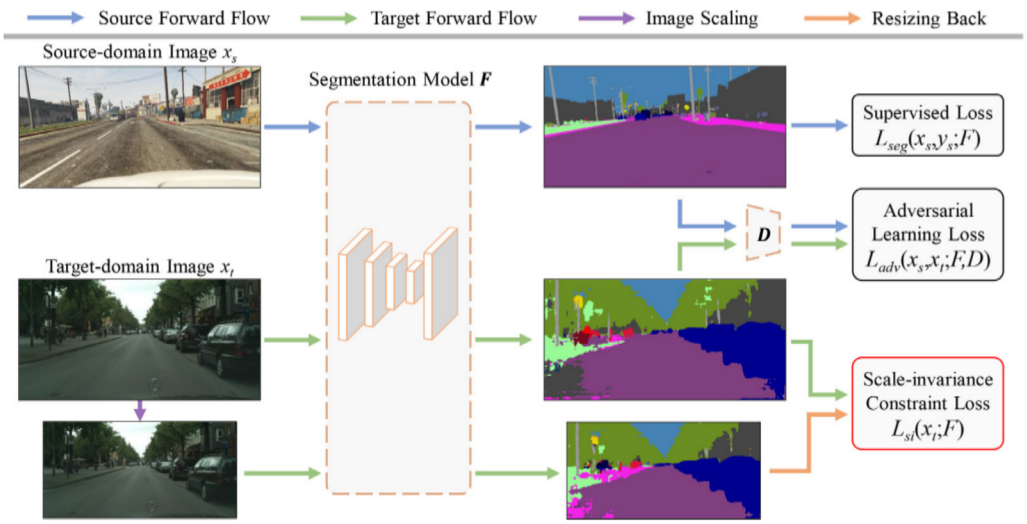

Fig. 2. The framework of the scale variance minimization (SVMin) in image segmentation: The classical inter-domain alignment via adversarial learning (with adversarial learning loss $\mathcal{L}_{adv}(x_s, x_t; F, D)$) may disintegrate the semantic structures of target-domain images and degrade the performance of image segmentation in the target domain. SVMin introduces a scale-invariance constraint (with scale-invariance constraint loss $\mathcal{L}_{si}(x_t; F)$) to preserve semantic structures of target-domain images while aligning target-domain feature representation to source-domain feature representation. It enforces the scale-invariance constraint by predicting two segmentation maps for each target-domain image and its scale transformation concurrently.

edge, this is the first work that leverages intra-domain invariance of target-domain images for UDA-based image segmentation.

3. Proposed methods

The proposed scale variance minimization (SVMin) method regulates the inter-domain alignment by enforcing the consistency of target-domain segmentation outputs across image scales as illustrated in Fig. 2. It also incorporates the idea of self-training to build SVMin based target-domain retraining for better image segmentation, more details to be described in the following subsections.

3.1. Problem definition

This work focuses on UDA in image segmentation. Given images $X_s \subset \mathbb{R}^{H\times W\times 3}$ with C-class pixel-level semantic labels $Y_s \subset (1,C)^{H\times W}$ in a source domain and unlabelled images $X_t \subset \mathbb{R}^{H\times W\times 3}$ in a target domain, the goal is to learn a image segmentation model F that performs well in the target domain. Inspired by the domain-divergence minimization [1], the prevalent approaches [15,28,30] address the UDA challenge by minimizing the discrepancy between the source and target domains. In the source domain, it trains a model F under a supervised loss $\mathcal{L}_{seg}$. In the target domain, F learns to extract domain-invariant features where a minimaxing game is played between F and a domain discriminator D under an adversarial learning loss $\mathcal{L}_{adv}$. The overall objective is a weighted combination of the two losses:

$$\mathcal{L}(F,D) = \mathcal{L}_{seg}(F) + \lambda_{adv}\mathcal{L}_{adv}(F,D). \quad (1)$$

where λ_{adv} is the weight that aims to balance the two losses.

Under the guidance of the adversarial loss, the minimization of the inter-domain discrepancy as defined in Eq. 1 strives to align the feature representation between the source and target domains. The brute-force inter-domain alignment may become negative when it alters the representation and disintegrates the semantic structures of many easily-segmented images in the target domain. We define this problem as a breach of consistency of semantic structures across scales, and propose a scale-invariance constraint to regulate the representation alignment across domains.

The proposed scale-invariance constraint is advantageous than the adversarial learning approach in domain adaptive image segmentation. The scale-invariance constraint is well-posed, i.e. the target of segmentation consistency across images of different scales is almost perfectly correct without considering the interpolation artefacts as introduced by image scaling. The network learning towards this objective introduces little side effect on image segmentation. On the contrary, the inter-domain alignment via adversarial learning is ill-posed as its objective is to align the feature distributions across domains. As the discrepancy across domains (e.g. the scene layout) exists inherently, the inter-domain alignment often crosses the line by disintegrating the semantic structures of target-domain images and leading to degraded segmentation.

Let $\mathcal{X}_s/\mathcal{X}_t$ denote the input distribution of source/target domain, and $\mathcal{Y}_s/\mathcal{Y}_t$ denote the labels of source/target samples. Along with the supervised training on the source domain (i.e., $F(\mathcal{X}_s) = \mathcal{Y}_s$), the objective is to find the optimal solution (i.e., $F(\mathcal{X}_t) = \mathcal{Y}_t$) by minimizing the inter-domain discrepancy. The adversarial learning approach minimizes the inter-domain discrepancy via brute-force inter-domain alignment at output space (i.e., $F(\mathcal{X}_t) = F(\mathcal{X}_s)$). This approach is sub-optimal in that target segmentation output $F(\mathcal{X}_t)$ may benefit from $\mathcal{Y}_t \approx \mathcal{Y}_s$, but suffer from $\mathcal{Y}_t \neq \mathcal{Y}_s$. As a result, this training objective may deconstruct the target-domain semantic structures consistency and destroy the optimal solution $F(\mathcal{X}_t) = \mathcal{Y}_t$ from the search space. To address this issue, the proposed scale-invariance constraint enforces intra-image semantic structure consistency (i.e., $\mathcal{R}^{-1}(F(\mathcal{R}(x_t))) = F(x_t)$), which will naturally lead target-domain semantic consistency. As the target of segmentation consistency across image scales is almost perfectly correct without considering the interpolation artefacts as introduced by image re-scaling, the scale-invariance constraint helps reduce the search space but keep the optimal solution (i.e., $F(\mathcal{X}_t) = \mathcal{Y}_t$) unaffected.

3.2. SVMin based inter-domain alignment

The proposed SVMin based inter-domain alignment aims to regulate inter-domain alignment by enforcing a scale-invariance constraint as illustrated in Fig. 2. In the source domain with pixel-level annotations, it feeds each original training image x_s to the segmentation network F that outputs a segmentation map $F(x_s)$ under a supervised segmentation loss $\mathcal{L}_{seg}$. In the target domain without pixel-level annotations, it feeds each original training image x_t and its scale transformation $\mathcal{R}(x_t)$ to the F that outputs two segmentation maps $F(x_t)$ and $F(\mathcal{R}(x_t))$ concurrently. Here, we em-

ployed bilinear interpolation for image scaling. The scaling ratio is randomly picked within certain ranges as described in the ensuing Implementation Details.

The adversarial learning aims to minimize the inter-domain discrepancy between $F(x_s)$ and $F(x_t)$ under an adversarial loss $\mathcal{L}_{adv}$. To regulate the inter-domain alignment, we introduce a scale-invariance constraint to minimize the discrepancy between $F(x_t)$ and $\mathcal{R}^{-1}(F(\mathcal{R}(x_t)))$ under a scale-invariance constraint loss $\mathcal{L}_{si}$.

Supervised loss In the source domain, we adopt the cross-entropy loss as the supervised loss $\mathcal{L}_{seg}$ to optimize the segmentation network F. Given a source-domain image $x_s \subset X_s$ and its corresponding labels $y_s \subset Y_s$, $\mathcal{L}_{seg}$ can be formulated as follows:

$$\mathcal{L}_{seg}(x_s, y_s; F) = \frac{1}{HW}\sum_{h,w}\sum_{c} -y_s^{(h,w,c)} \log F(x_s)^{(h,w,c)}. \tag{2}$$

Note we did not implement the scale-invariance constraint and compute scale-invariance loss in the source domain as the strong pixel-level supervision helps to keep the semantic image structures well.

Adversarial learning loss We introduce a discrimination network D to align the feature representation and minimize the discrepancy between the source and target domains as illustrated in Fig. 2. Given a source-domain image $x_s \subset X_s$ and a target-domain image $x_t \subset X_t$, the adversarial learning loss $\mathcal{L}_{adv}$ can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{adv}(x_s, x_t; F, D) = &\log(D(-F(x_s)\log F(x_s))) \\ &+ \log(1 - D(-F(x_t)\log F(x_t))).\end{aligned} \tag{3}$$

Scale-invariance constraint loss We design a scale-invariance constraint loss to enforce the consistency of target-domain semantic structures and regulate the adversarial learning loss in inter-domain alignment. Given a target-domain image $x_t \subset X_t$, the scale-invariance constraint loss $\mathcal{L}_{si}$ is formulated as:

$$\mathcal{L}_{si}(x_t; F) = \frac{1}{HWC}\sum_{h,w,c} |F(x_t)^{(h,w,c)} - \mathcal{R}^{-1}(F(\mathcal{R}(x_t)^{(h,w,c)}))|. \tag{4}$$

Training objective The objective function of the proposed SVMin based adversarial learning model (SVMin_AL) can thus be formulated by summing up the three training losses as follows:

$$\mathcal{L}_{\text{SVMin\_AL}}(F, D) = \mathcal{L}_{seg}(F) + \lambda_{adv}\mathcal{L}_{adv}(F, D) + \lambda_{si}\mathcal{L}_{si}(F). \tag{5}$$

where λ_{si} is the weight that aims to balance the supervised loss and scale-invariance constraint loss.

The optimization of the SVMin model $F_{\text{SVMin\_AL}}$ can be formulated as:

$$F_{\text{SVMin\_AL}} = \arg\min_{F}\max_{D} \mathcal{L}_{\text{SVMin\_AL}}(F, D). \tag{6}$$

3.3. SVMin based target-domain retraining

Self-training has been widely explored as an effective fine-tuning strategy in domain adaptive image segmentation [13,39]. It works by generating pseudo labels $\hat{Y}_t$ from confident predictions in the target domain. Current state-of-the-art approaches [9,32,35] explore intra-domain knowledge via retraining in target domain with generated pseudo labels. We fine-tune $F_{\text{SVMin\_AL}}$ by including a target-domain retraining loss beyond the scale-invariance loss to explore more knowledge in the target domain. Specifically, we employ the scale-invariance loss to enhance the consistency of target-domain semantic structures during the retraining process. Given a target-domain image $x_t \subset X_s$ and the generated pseudo labels $\hat{y}_t \subset \hat{Y}_t$ from [39], the target-domain retraining loss can be formulated as follows:

$$\mathcal{L}_{rt}(x_t, \hat{y}_t; F) = \tfrac{1}{HW}\textstyle\sum_{h,w}\sum_{c} -\hat{y}_t^{(h,w,c)} \log F(x_t^{(h,w,c)}) \tag{7}$$

Combining the target-domain retraining loss and scale-invariance loss, the $F_{\text{SVMin\_AL}}$ can be fine-tuned to obtain final model as follows:

$$\mathcal{L}_{\text{SVMin\_AL\_TR}} = \arg\min_{F_{\text{SVMin\_AL}}} (\mathcal{L}_{idst}(F_{\text{SVMin\_AL}}) + \mathcal{L}_{si}(F_{\text{SVMin\_AL}})). \tag{8}$$

3.4. Theoretical insights

The scale-invariance constraint is inherently connected with a contemporary UDA theory. It is actually an example of domain-divergence minimization [1]:

Proposition. *The SVMin can be modelled as a cross-domain $\mathcal{H}$-divergence minimization problem that can be optimized with L1-normalization.*

Proof: The scale-invariance constraint is inherently connected with the contemporary domain adaptation theory [1]. Consider a segmentation model F in hypothesis space $\mathcal{H}_F$ over the output-space representation $F(x)$ on source domain input distribution $\mathcal{X}_s$ and target domain input distribution $\mathcal{X}_t$. Let $\epsilon_{\mathcal{X}_s}(F)$ denote the error of a hypothesis $F \in \mathcal{H}_F$ for $\mathcal{X}_s$, $\epsilon_{\mathcal{X}_s,\mathcal{X}_t}$ denote the constant error of the single ideal hypothesis for both $\mathcal{X}_s$ and $\mathcal{X}_t$, and $d_{\mathcal{H}}(\mathcal{X}_s, \mathcal{X}_t)$ denote the cross-domain $\mathcal{H}$-divergence between $\mathcal{X}_s$ and $\mathcal{X}_t$. The target error $\epsilon_{\mathcal{X}_t}(F)$ of hypothesis F is bounded by Ben-David et al. [1]:

$$\epsilon_{\mathcal{X}_t}(F) \le \epsilon_{\mathcal{X}_s}(F) + \epsilon_{\mathcal{X}_s,\mathcal{X}_t} + d_{\mathcal{H}}(\mathcal{X}_s, \mathcal{X}_t). \tag{9}$$

Because the first and second terms correspond to the well-studied supervised learning problems, the goal of UDA is to reduce the third term, i.e., the domain divergence:

$$d_{\mathcal{H}}(\mathcal{X}_s, \mathcal{X}_t) \triangleq 2\sup_{h\in\mathcal{H}} |\Pr_{x_s\sim\mathcal{X}_s}[h(F(x_s)) = 1] - \Pr_{x_t\sim\mathcal{X}_t}[h(F(x_t)) = 1]|. \tag{10}$$

We measure the divergence between two distributions $\mathcal{X}_s$ and $\mathcal{X}_t$ based on the segmentation variance (SV) at different image scales. Let h be a domain classifier that decides the binary domain label of x by the value of its segmentation variance (SV) at different image scales, namely,

$$h(F(x)) = \begin{cases} 1, & \text{if } SV(x) \ge \xi, \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where $SV(x) = |F(x) - \mathcal{R}^{-1}(F(\mathcal{R}(x)))|$ is the segmentation variance at different image scales and ξ is a small threshold to determine the domain label.

Given these equations, we show that the scale-invariance constraint is related to $d_{\mathcal{H}}(\mathcal{X}_s, \mathcal{X}_t)$, where Eq. (10) can be rewritten as:

$$\begin{aligned} d_{\mathcal{H}}(\mathcal{X}_s, \mathcal{X}_t) &\triangleq 2\sup_{h\in\mathcal{H}} |\Pr_{x_s\sim\mathcal{X}_s}[h(F(x_s)) = 1] - \Pr_{x_t\sim\mathcal{X}_t}[h(F(x_t)) = 1]| \\ &= 2\sup_{F\in\mathcal{H}_F} |\Pr_{x_s\sim\mathcal{X}_s}[SV(x_s) \ge \xi] - \Pr_{x_t\sim\mathcal{X}_t}[SV(x_t) \ge \xi]| \\ &\le 2\sup_{F\in\mathcal{H}_F} \Pr_{x_t\sim\mathcal{X}_t}[SV(x_t) \ge \xi]. \end{aligned} \tag{12}$$

The last inequality in Eq. (12) holds because the segmentation variance (SV) at different image scales is very small on source-domain samples, where there are sufficient labelled training data for minimizing supervised losses. Therefore, the objective is to learn a model F to achieve lowest upper bound of cross-domain $\mathcal{H}$-divergence:

$$\min_{F\in\mathcal{H}_F} \Pr_{x_t\sim\mathcal{X}_t}[|F(x_t) - \mathcal{R}^{-1}(F(\mathcal{R}(x_t)))|] \ge \xi]. \tag{13}$$

Thus, the proposed SVMin can be modeled as a cross-domain $\mathcal{H}$-divergence minimization problem that can be optimized with L1-normalization.

Table 1
Comparisons of our method (SVMin_AL_TR) with the state-of-the-art over GTA5 → Cityscapes: our method outperforms state-of-the-art clearly. *V* and *R* refer to VGG16 and ResNet101 backbones.

GTA5 → Cityscapes																					
Methods	model	road	side.	build.	wall	fence	pole	light	sign	vege.	terr.	sky	pers.	rider	car	truck	bus	train	motor	bike	mIoU
AdaptSeg [28]	V	87.3	29.8	78.6	21.1	18.2	22.5	21.5	11.0	79.7	29.6	71.3	46.8	6.5	80.1	23.0	26.9	0.0	10.6	0.3	35.0
AdvEnt [30]	V	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
CLAN [15]	V	88.0	30.6	79.2	23.4	20.5	26.1	23.0	14.8	81.6	34.5	72.0	45.8	7.9	80.5	**26.6**	29.9	0.0	10.7	0.0	36.6
CrCDA [8]	V	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
BDL [13]	V	89.2	40.9	81.2	29.1	19.2	14.2	29.0	19.6	83.7	35.9	80.7	54.7	**23.3**	82.7	25.8	28.0	2.3	**25.7**	19.9	41.3
FDA [35]	V	86.1	35.1	80.6	**30.8**	20.4	27.5	30.0	26.0	82.1	30.3	73.6	52.5	21.7	81.7	24.0	30.5	**29.9**	14.6	**24.0**	42.2
SIM [32]	V	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
Ours	V	**89.7**	**42.1**	82.6	29.3	22.5	**32.3**	**35.5**	**32.2**	**84.6**	35.4	77.2	**61.6**	21.9	**86.2**	26.1	36.7	7.7	16.9	19.4	**44.2**
AdaptSeg [28]	R	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
CLAN [15]	R	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2
AdvEnt [30]	R	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	32.4	45.5
IDA [18]	R	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
PatAlign [29]	R	92.3	51.9	82.1	29.2	25.1	24.5	33.8	33.0	82.4	32.8	82.2	58.6	27.2	84.3	33.4	46.3	2.2	29.5	32.3	46.5
CRST [39]	R	91.0	55.4	80.0	33.7	21.4	37.3	32.9	24.5	85.0	34.1	80.8	57.7	24.6	84.1	27.8	30.1	**26.9**	26.0	42.3	47.1
BDL [13]	R	91.0	44.7	84.2	**34.6**	27.6	30.2	36.0	36.0	85.0	**43.6**	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5
CrCDA [8]	R	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
SIM [32]	R	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
CAG [37]	R	90.4	51.6	83.8	34.2	27.8	38.4	25.3	48.4	85.4	38.2	78.1	58.6	**34.6**	84.7	21.9	42.7	41.1	29.3	37.2	50.2
TIR [9]	R	**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
FDA [35]	R	92.5	53.3	82.4	26.5	27.6	36.4	40.6	38.9	82.3	39.8	78.0	62.6	34.4	84.9	34.1	**53.1**	16.9	27.7	**46.4**	50.5
Ours	R	**92.9**	**56.2**	84.3	34.0	22.0	**43.1**	**50.9**	**48.6**	**85.8**	42.0	78.9	**66.6**	26.9	**88.4**	35.2	46.0	10.9	25.4	39.6	**51.5**

4. Experiments

4.1. Datasets

We evaluate our approach over two challenging UDA-based image segmentation tasks: GTA5 → Cityscapes and SYNTHIA → Cityscapes. To demonstrate the genericity of our approach, we perform preliminary evaluations over a UDA-based image classification task. The three tasks involves four datasets as listed.

Cityscapes [3] as the target-domain dataset in our experiment has been widely used in image segmentation. It consists of 2975 real-world images for training and 500 for evaluation.

GTA5 [21] contains 24,966 high-resolution synthesized images. It shares 19 common pixel classes with the Cityscapes, the target-domain dataset as used in our experiments.

SYNTHIA [22] contains 9400 synthetic images. It shares 16 common pixel classes with the Cityscapes as used in our experiments [28,30,35].

VisDA [19] is a domain adaptive image classification dataset. It consists of a training set with synthetic renderings of 3D models, a validation and test set with real-world images.

4.2. Implementation details

Similar to [28,30,35], we use Deeplab-V2 architecture [2] as the segmentation network F and apply atrous spatial pyramid pooling to the extracted feature maps with sampling rate fixed at 6; 12; 18; 24. DeepLab-v2 [2] is built on a deep convolutional neural network for feature map extraction. It employs dilated convolution in the last two convolution layers to enlarge the receptive field. In addition, it introduces atrous spatial pyramid pooling to capture image context information by using multi-scale parallel filters. For the discriminator network D, we follow [28,30,32] and employ 5 convolution layers with kernel size 4×4, stride of 2×2, and channel numbers of {64, 128, 256, 512, 1} for each layer. During training, we utilize classical stochastic gradient descent algorithm to optimize our networks with a momentum of 0.9 and a weight decay of $1e-4$. The initial learning rate is set at $2.5e-4$ and decayed by a polynomial policy with a power of 0.9 as illustrated in Chen et al. [2]. The balancing weight λ_{adv} is empirically set to 0.001 following [28,30] and λ_{si} is set to 1. The scaling ratio is randomly selected in the range [0.8,1.2]. For the comparison with state-of-the-art methods, we evaluate two network backbones including VGG16 [27] and ResNet101 [6] both pre-trained on ImageNet. All experiments are implemented on a single Tesla V100 GPU by employing PyTorch toolbox, and the maximum memory usage is 12 GB.

4.3. Comparison with state-of-art

We evaluate our approach over two widely studied domain adaptive image segmentation tasks GTA5→Cityscapes and SYNTHIA→Cityscapes, and compared it with a number of state-of-the-art methods as shown in Tables 1 and 2. We adopted two network backbones ResNet101 and VGG16 and the widely used evaluation metric, i.e., mean intersection over union (mIoU), in evaluations. As the two tables show, our approach outperforms all state-of-the-art methods clearly and consistently across the two tasks and the two network backbones. The superior segmentation performance is largely thanks to the proposed scale-invariance constraint that introduces certain supervision and endeavours to keep the integrity of image semantic structures while aligning the representation of target domain to that of source domain. Without the scale-invariance constraint, state-of-the-art methods may over-align representation which disintegrates the semantic structures and degrades the segmentation of images in the target domain.

We provide qualitative results in the GTA5 → Cityscapes task in Fig. 3. It is obvious that segmentation inconsistency maps (segmentation variance across different image scales) can well reflect the image segmentation performance, where the segmentation of regions with high segmentation inconsistency (darker) is normally bad and noisy while that of regions with low segmentation inconsistency (lighter) is normally good and clear. In other words, the segmentation inconsistency map can actually serve as a good indicator to show how well the segmentation is. Our proposed approach utilizes this property and minimizes segmentation inconsistency effectively, leading to superior segmentation performances as illustrated in Fig. 3. As a comparison, state-of-the-art UDA approaches CRST [39], TIR [9] and FDA [35] produce very promising segmentation, but they cannot segment target images consistently across different image scales. Specifically, the segmentation around high variance regions is clearly noisy with various segmen-

Table 2
Comparisons of our method (SVMin_AL_TR) with the state-of-the-art over the task SYNTHIA → Cityscapes: our method surpasses state-of-the-art by large margins. *V* and *R* refer to VGG16 and ResNet101 backbones. *mIoU* and *mIoU** are computed over 16 and 13 pixel classes, respectively.

SYNTHIA → Cityscapes																			
Methods	model	road	side.	build.	wall	fence	pole	light	sign	vege.	sky	pers.	rider	car	bus	motor	bike	mIoU	mIoU*
AdaptSeg [28]	V	78.9	29.2	75.5	–	–	–	0.1	4.8	72.6	76.7	43.4	8.8	71.1	16.0	3.6	8.4	–	37.6
AdvEnt [30]	V	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	36.6
CLAN [15]	V	80.4	30.7	74.7	–	–	–	1.4	8.0	77.1	79.0	46.5	8.9	73.8	18.2	2.2	9.9	–	39.3
CrCDA [8]	V	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	41.1
BDL [13]	V	72.0	30.3	74.5	0.1	0.3	24.6	10.2	25.2	**80.5**	80.0	54.7	**23.2**	72.7	24.0	7.5	44.9	39.0	46.1
FDA [35]	V	**84.2**	**35.1**	**78.0**	6.1	0.4	**27.0**	8.5	22.1	77.2	79.6	55.5	19.9	74.8	**24.9**	**14.3**	40.7	40.5	47.3
Ours	V	82.5	31.5	77.6	**7.6**	**0.7**	26.0	**12.3**	**28.4**	79.4	**82.1**	**58.9**	21.5	**82.1**	22.1	9.6	**49.2**	**41.9**	**49.0**
PatAlign [29]	R	82.4	38.0	78.6	8.7	0.6	26.0	3.9	11.1	75.5	84.6	53.5	21.6	71.4	32.6	19.3	31.7	40.0	46.5
AdaptSeg [28]	R	84.3	42.7	77.5	–	–	–	4.7	7.0	77.9	82.5	54.3	21.0	72.3	32.2	18.9	32.3	–	46.7
CLAN [15]	R	81.3	37.0	80.1	–	–	–	16.1	13.7	78.2	81.5	53.4	21.2	73.0	32.9	22.6	30.7	–	47.8
AdvEnt [30]	R	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	48.0
IDA [18]	R	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	48.9
TIR [9]	R	**92.6**	**53.2**	79.2	–	–	–	1.6	7.5	78.6	84.4	52.6	20.0	82.1	34.8	14.6	39.4	–	49.3
CrCDA [8]	R	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	50.0
CRST [39]	R	67.7	32.2	73.9	10.7	**1.6**	**37.4**	22.2	**31.2**	80.8	80.5	60.8	29.1	82.8	25.0	19.4	45.3	43.8	50.1
BDL [13]	R	86.0	46.7	80.3	–	–	–	14.1	11.6	79.2	81.3	54.1	27.9	73.7	**42.2**	25.7	45.3	–	51.4
SIM [32]	R	83.0	44.0	80.3	–	–	–	17.1	15.8	80.5	81.8	59.9	**33.1**	70.2	37.3	28.5	45.8	–	52.1
FDA [35]	R	79.3	35.0	73.2	–	–	–	19.9	24.0	61.7	82.6	61.4	31.1	83.9	40.8	**38.4**	**51.1**	–	52.5
CAG [37]	R	84.7	40.8	81.7	7.8	0.0	35.1	13.3	22.7	84.5	77.6	64.2	27.8	80.9	19.7	22.7	48.3	44.5	52.6
Ours	R	89.8	47.7	**82.3**	**14.4**	0.2	37.1	**35.4**	22.1	**85.1**	**84.9**	**65.8**	25.6	**86.0**	30.5	31.0	50.7	**49.3**	**56.7**

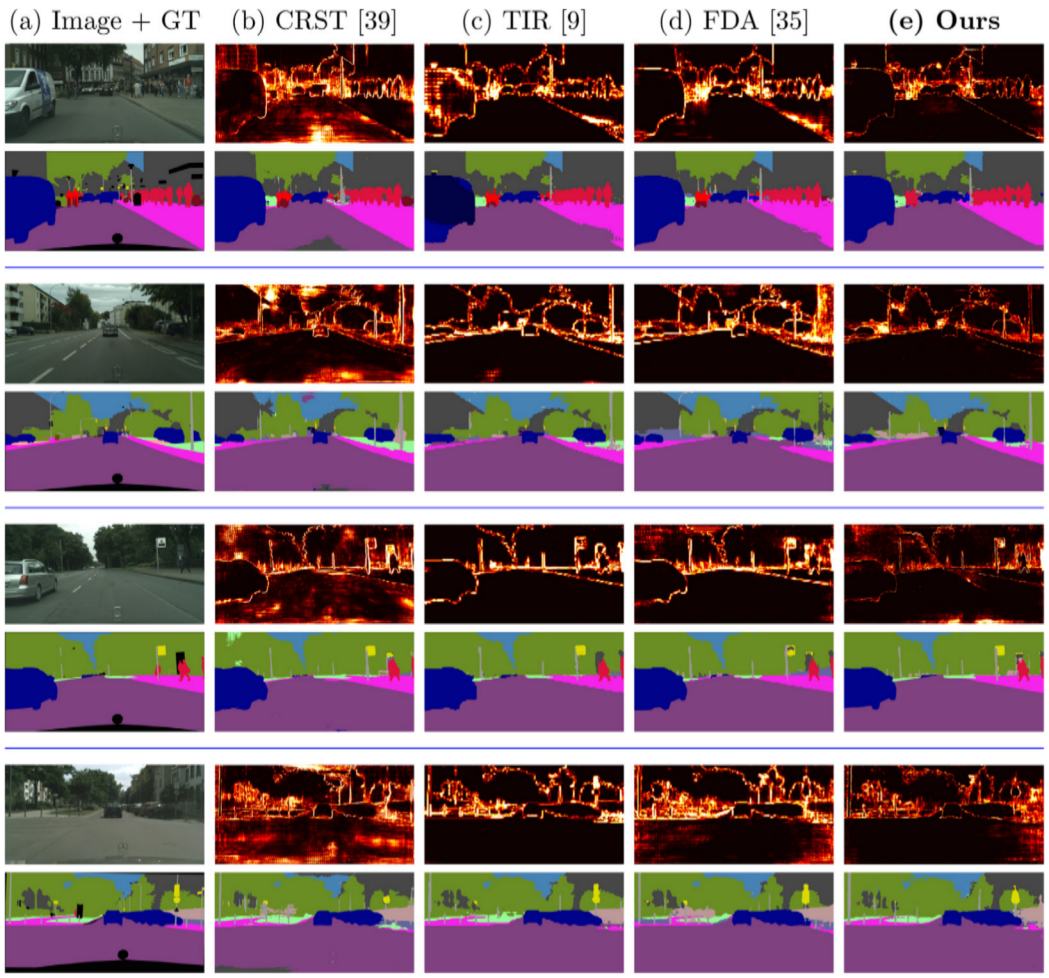

Fig. 3. Qualitative segmentation for GTA5 → Cityscapes: Column (a) shows target-domain sample images at the top and the corresponding ground-truth segmentation at the bottom. Column (b)–(e) show the segmentation (at the bottom) as produced by state-of-the-art UDA approaches CRST [39], TIR [9], FDA [35] and our proposed method (SVMin_AL_TR), as well as the corresponding inconsistency maps as computed between the segmentation of original-scale images and re-scaled images (at the top), respectively. Best viewed in colors.

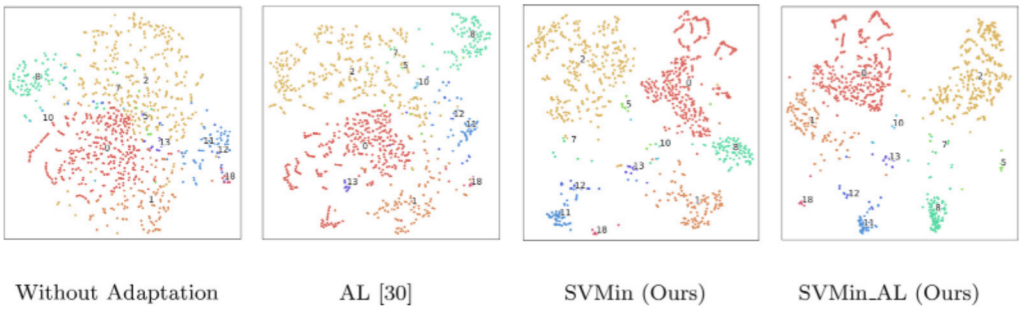

Fig. 4. t-SNE [16] visualization of feature distribution for Cityscapes images in the target domain: Each color represents one specific semantic classes of image pixels with a digit showing the class centre. σ_w^2 and σ_b^2 on the top of each graph are intra-class variance and inter-class distance of the corresponding feature distribution. Our approach outperforms the prevalent adversarial learning (AL) [30] in UDA qualitatively and quantitatively.

Table 3
Ablation study of our approach over GTA5 → Cityscapes task: SVMin outperforms adversarial learning (AL) clearly, and the two approaches are complementary (using ResNet101 backbone). SVMin_AL based target-domain retraining (SVMin_AL_TR) further improves the target-domain retraining (TR) method in model fine-tuning.

	Inter-domain alignment			target-domain retraining		mIoU
Method	L_{seg}	L_{adv}	L_{si}	L_{rt}	L_{si}	
Baseline	✓					36.6
AL [30]	✓	✓				43.8
SVMin	✓		✓			46.5
SVMin_AL	✓	✓	✓			48.1
SVMin_AL + TR	✓	✓	✓	✓		50.6
SVMin_AL_TR	✓	✓	✓	✓	✓	**51.5**

tation errors. In contrast, our approach segments target images with lower segmentation variance, leading to better and smooth segmentation output. Besides, most segmentation variance in our approach locate at the category transition areas, which are naturally very difficult to correctly or consistently segment, even for fully supervised models.

4.4. Ablation studies

We perform extensive ablation studies to demonstrate how the scale-invariance constraint improves UDA and complements adversarial learning in inter-domain alignment. As listed in Table 3, we trained 6 models for the task GTA5→Cityscapes including (1) **Baseline** that is trained using *Supervised Loss* $L_{seg}(x_s, y_s; F)$ only with no adaptation, (2) **AL** [30] that is trained using *Adversarial Learning Loss* $L_{adv}(x_s, x_t; F, D)$ and $L_{seg}(x_s, y_s; F)$ only, (3) **SVMin** that is trained using *Scale-Invariance Constrain Loss* $L_{si}(x_t; F)$ and $L_{seg}(x_s, y_s; F)$ only, (4) **SVMin_AL** that is trained using all three losses $L_{seg}(x_s, y_s; F)$, $L_{adv}(x_s, x_t; F, D)$ and $L_{si}(x_t; F)$, (5) **SVMin_AL+TR** that fine-tunes the SVMin_AL model using the target-domain retraining (TR) loss $\mathcal{L}_{rt}(x_t, \hat{y}_t; F)$, and (6) **SVMin_AL_TR** that fine-tunes the SVMin_AL model via combining $\mathcal{L}_{rt}(x_t, \hat{y}_t; F)$ and $L_{si}(x_t; F)$.

We applied the six models to conduct evaluation and Table 3 shows experimental results. It can be seen that the **Baseline** does not performs well due to the discrepancy across domains. In addition, both **AL** and **SVMin** outperform the **Baseline** by large margins, demonstrating the importance of UDA for minimizing domain discrepancy. Nevertheless, **SVMin** outperforms **AL** clearly. This shows that scale-invariance constraint (for target-domain consistency) is more effective than adversarial learning (for inter-domain alignment) in UDA-based image segmentation. Further, **SVMin_AL** performs clearly the best, demonstrating that the two orthogonal UDA approaches are actually complementary to each other. Finally, **SVMin_AL_TR** outperforms **SVMin_AL+TR**,

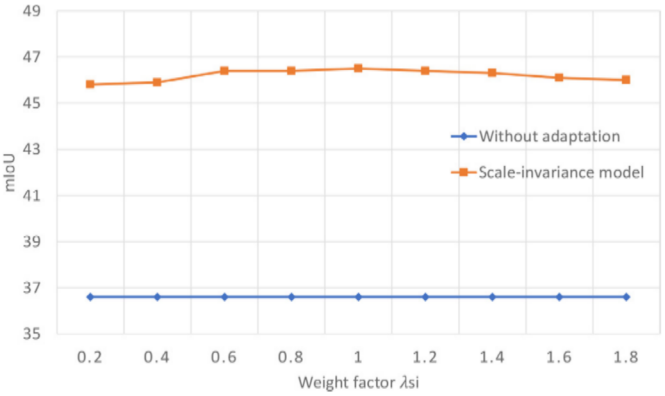

Fig. 5. Parameter learning of weight factor λ_{si} in Eq. (5) over GTA5 → Cityscapes: The segmentation performance is robust to various weight factors, which further demonstrates the scale-invariance constraint loss is a well-posed objective (using ResNet101 backbone).

Table 4
The scale-invariance constraint loss L_{si} can be easily incorporated into existing UDA networks as a plug-in with consistent performance improvement.

	GTA5 → Cityscapes			SYNTHIA → Cityscapes		
Method	base	$+L_{si}$	gain	base	$+L_{si}$	gain
Adapt-SegMap [28]	42.4	47.1	+4.7	46.7	50.5	+3.8
MinEnt [30]	42.3	46.9	+4.6	44.2	49.4	+5.2
CLAN [15]	43.2	47.7	+4.5	47.8	51.7	+3.9
AdvEnt [30]	43.8	48.1	+4.3	47.6	51.8	+4.2

demonstrating the importance of the scale-invariance constraint in model fine-tuning in the target domain.

For the Cityscapes images in the target domain, Fig. 4 shows the feature distributions of 19 semantic pixel classes that are produced by the first 4 ablation study models from left to right, respectively. As Fig. 4 shows, both qualitative t-SNE visualization and quantitative intra-class variance and inter-class distances are well aligned with the image segmentation in Table 3.

4.5. Parameter learning

We study the impact of the weight factor λ_{si} in Eq. (5) that is used to control the scale invariance constraint. Fig. 5 shows that the segmentation performance is robust to λ_{si} in SVMin model (i.e., $\lambda_{si} = 0.2 \sim 1.8 \rightarrow mIoU = 45.8\ \% \sim 46.5\ \%$). These results further demonstrate that the scale-invariance constraint loss is a well-posed objective as it is almost perfectly correct without considering the interpolation artefacts as introduced by image rescaling. On the contrary, the segmentation performance is very unstable to the weight factor of adversarial learning loss (λ_{adv}) around

Table 5
Comparison of our method (SVMin_AL_TR) with the state-of-the-art in domain adaptive image classification over VisDA17: our method achieved superior classification over both validation data and test data.

Method	data	aero	bike	bus	car	horse	knife	motor	person	plant	skate.	train	truck	mean
CDAN [31]	Val	–	–	–	–	–	–	–	–	–	–	–	–	71.4
MCD [25]	Val	87	60.9	83.7	64	88.9	79.6	84.7	76.9	88.6	40.3	83	25.8	71.9
ADR [24]	Val	87.8	79.5	83.7	65.3	92.3	61.8	88.9	73.2	87.8	60	85.5	32.3	74.8
SAFN [34]	Val	93.6	61.3	84.1	70.6	**94.1**	79.0	**91.8**	79.6	89.9	55.6	**89.0**	24.2	76.1
SWD [10]	Val	90.8	**82.5**	81.7	70.5	91.7	69.5	86.3	77.5	87.4	63.6	85.6	29.2	76.4
GTA [26]	Val	–	–	–	–	–	–	–	–	–	–	–	–	77.1
CRST [39]	Val	88.0	79.2	61.0	60.0	87.5	81.4	86.3	78.8	85.6	**86.6**	73.9	**68.8**	78.1
Ours	Val	94.0	70.9	**86.1**	**75.4**	94.0	**89.5**	90.2	**82.1**	**93.1**	79.0	82.5	25.7	**80.2**
GTA [26]	Test	–	–	–	–	–	–	–	–	–	–	–	–	72.3
Ours	Test	89.5	46.5	86.6	92.9	83.9	72.7	76.9	68.6	95.9	60.3	74.7	46.2	**74.6**

the small empirical value 0.001 [28,30] in AL model, (i.e., $\lambda_{adv} = 0.0002 \sim 0.0018 \rightarrow mIoU = 39.6\ \% \sim 43.8\ \%$). This is largely because the inter-domain alignment via adversarial learning is ill-posed as its objective is to align the feature space across domains.

4.6. Discussion

The proposed scale-invariance constraint has two unique features in UDA. First, it can be easily incorporated into existing UDA methods as a plug-in with consistent performance boost but little extra parameters and computation. We evaluate this feature over UDA-based image segmentation task by including the scale-invariance constraint into several UDA methods as listed in Table 4. We can see that the inclusion of the *scale-invariance constraint loss* L_{si} improves the image segmentation of state-of-the-art methods consistently for both GTA5 $\rightarrow$ Cityscapes and SYNTHIA $\rightarrow$ Cityscapes. As the inclusion of L_{si} has little effect over the network structures, the inference has little extra parameters and computation once the model is trained.

Second, the proposed scale-invariance constraint is generic and can apply to other tasks with slight adaptation. We validate this feature by conducting a preliminary UDA-based image classification experiment over VisDA [19]. Since the original and scaled images produce class probability vectors of the same size, the *scale-invariance constraint loss* $\mathcal{L}_{si}$ is re-defined as follows:

$$\mathcal{L}_{si}(x_t; F) = \frac{1}{C}\sum_{c} |F(x_t)^{(c)} - F(\mathcal{R}(x_t))^{(c)}|. \tag{14}$$

In the experiment, we used the VisDA training set as the source domain and the validation set as the target domain as in Lee et al. [10], Sankaranarayanan et al. [26], Zou et al. [39]. Additionally, we also evaluate the model on the VisDA test set following [26] as shown in Table 5. It is clear that SVMin outperforms state-of-the-art CRST [39] by 2.1 on the validation set and GTA [26] by 2.3 on the test set, demonstrating its genericity in different domain adaptive tasks.

5. Conclusions

This paper presents a scale variance minimization (SVMin) technique that exploits a scale-invariance constraint as certain supervision to improve domain adaptive semantic segmentation performance. SVMin leverages an intrinsic property of images that simple scale transformation has little effect on the semantic structures of images. It enforces the semantic structures of images by imposing consistency over the segmentation of an image and its scale transformation, hence achieving superior semantic segmentation performance. In addition, the scale-invariance constraint is generic and can be extended to other domain adaptive computer vision tasks such as image classification.

On the other hand, the proposed scale-invariance is essentially an intra-image constraint. It exploits the semantic consistency among the segmentation of the same image of different scales, but ignores the discrepancy across domains. It is therefore orthogonal and complementary to the adversarial learning based domain adaptation and alignment techniques. In the future work, we will investigate how to optimally fuse the proposed scale invariance constraint and domain alignment technique so that the domain adaptation can leverage both intra-domain and inter-domain information concurrently. In addition, we will study how to extend SVMin to transform unsupervised domain adaption to semi-supervised/weak-supervised domain adaptation so as to achieve comparable performance with fully-supervised learning.

Declaration of Competing Interest

Authors declare that they have no conflict of interest.

Acknowledgments

This research was conducted at Singtel Cognitive and Artificial Intelligence Lab for Enterprises (SCALE@NTU), which is a collaboration between Singapore Telecommunications Limited (Singtel) and Nanyang Technological University (NTU) that is funded by the Singapore Government through the Industry Alignment Fund - Industry Collaboration Projects Grant.

References

[1] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira, Analysis of representations for domain adaptation, in: Advances in Neural Information Processing Systems, 2007, pp. 137–144.
[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.
[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.
[4] C. Deng, X. Liu, C. Li, D. Tao, Active multi-kernel domain adaptation for hyperspectral image classification, Pattern Recognit. 77 (2018) 306–315.
[5] M. Everingham, S.M.A. Eslami, L.V. Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: aretrospective, Int. J. Comput. Vis. 111 (1) (2015) 98–136.
[6] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[7] W. Hong, Z. Wang, M. Yang, J. Yuan, Conditional generative adversarial network for structured domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1335–1344.
[8] J. Huang, S. Lu, D. Guan, X. Zhang, Contextual-relation consistent domain adaptation for semantic segmentation, in: European Conference on Computer Vision, Springer, 2020, pp. 705–722.
[9] M. Kim, H. Byun, Learning texture invariant representation for domain adaptation of semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12975–12984.
[10] C.-Y. Lee, T. Batra, M.H. Baig, D. Ulbricht, Sliced Wasserstein discrepancy for unsupervised domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10285–10295.

[11] R. Li, W. Cao, Q. Jiao, S. Wu, H.-S. Wong, Simplified unsupervised image translation for semantic segmentation adaptation, Pattern Recognit. 105 (2020) 107343.
[12] Y. Li, N. Wang, J. Shi, X. Hou, J. Liu, Adaptive batch normalization for practical domain adaptation, Pattern Recognit. 80 (2018) 109–117.
[13] Y. Li, L. Yuan, N. Vasconcelos, Bidirectional learning for domain adaptation of semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6936–6945.
[14] J. Liang, R. He, Z. Sun, T. Tan, Exploring uncertainty in pseudo-label guided unsupervised domain adaptation, Pattern Recognit. 96 (2019) 106996.
[15] Y. Luo, L. Zheng, T. Guan, J. Yu, Y. Yang, Taking a closer look at domain shift: category-level adversaries for semantics consistent domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2507–2516.
[16] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605. (Nov).
[17] K. Nguyen, C. Fookes, S. Sridharan, Context from within: hierarchical context modeling for semantic segmentation, Pattern Recognit. 105 (2020) 107358.
[18] F. Pan, I. Shin, F. Rameau, S. Lee, I.S. Kweon, Unsupervised intra-domain adaptation for semantic segmentation through self-supervision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3764–3773.
[19] X. Peng, B. Usman, N. Kaushik, D. Wang, J. Hoffman, K. Saenko, VisDA: a synthetic-to-real benchmark for visual domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 2021–2026.
[20] M.M. Rahman, C. Fookes, M. Baktashmotlagh, S. Sridharan, Correlation-aware adversarial domain adaptation and generalization, Pattern Recognit. 100 (2020) 107124.
[21] S.R. Richter, V. Vineet, S. Roth, V. Koltun, Playing for data: ground truth from computer games, in: European Conference on Computer Vision, Springer, 2016, pp. 102–118.
[22] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, A.M. Lopez, The SYNTHIA dataset: a large collection of synthetic images for semantic segmentation of urban scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3234–3243.
[23] K. Saito, Y. Ushiku, T. Harada, Asymmetric tri-training for unsupervised domain adaptation, in: International Conference on Machine Learning, 2017, pp. 2988–2997.
[24] K. Saito, Y. Ushiku, T. Harada, K. Saenko, Adversarial dropout regularization, in: International Conference on Learning Representations, 2018.
[25] K. Saito, K. Watanabe, Y. Ushiku, T. Harada, Maximum classifier discrepancy for unsupervised domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3723–3732.
[26] S. Sankaranarayanan, Y. Balaji, C.D. Castillo, R. Chellappa, Generate to adapt: Aligning domains using generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8503–8512.
[27] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Machine Learning, 2015.
[28] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, M. Chandraker, Learning to adapt structured output space for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7472–7481.
[29] Y.-H. Tsai, K. Sohn, S. Schulter, M. Chandraker, Domain adaptation for structured output via discriminative patch representations, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1456–1465.
[30] T.-H. Vu, H. Jain, M. Bucher, M. Cord, P. Pérez, Advent: adversarial entropy minimization for domain adaptation in semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2517–2526.
[31] X. Wang, Y. Jin, M. Long, J. Wang, M.I. Jordan, Transferable normalization: towards improving transferability of deep neural networks, in: Advances in Neural Information Processing Systems, 2019, pp. 1951–1961.
[32] Z. Wang, M. Yu, Y. Wei, R. Feris, J. Xiong, W.-m. Hwu, T.S. Huang, H. Shi, Differential treatment for stuff and things: a simple unsupervised domain adaptation method for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12635–12644.
[33] C.-D. Xu, X.-R. Zhao, X. Jin, X.-S. Wei, Exploring categorical regularization for domain adaptive object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11724–11733.
[34] R. Xu, G. Li, J. Yang, L. Lin, Larger norm more transferable: an adaptive feature norm approach for unsupervised domain adaptation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1426–1435.
[35] Y. Yang, S. Soatto, FDA: fourier domain adaptation for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4085–4095.
[36] P. Zhang, W. Liu, H. Wang, Y. Lei, H. Lu, Deep gated attention networks for large-scale street-level scene segmentation, Pattern Recognit. 88 (2019) 702–714.
[37] Q. Zhang, J. Zhang, W. Liu, D. Tao, Category anchor-guided unsupervised domain adaptation for semantic segmentation, in: Advances in Neural Information Processing Systems, 2019, pp. 433–443.
[38] H. Zhu, F. Meng, J. Cai, S. Lu, Beyond pixels: a comprehensive survey from bottom-up to semantic image segmentation and cosegmentation, J. Vis. Commun. Image Represent. 34 (2016) 12–27.
[39] Y. Zou, Z. Yu, X. Liu, B. Kumar, J. Wang, Confidence regularized self-training, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5982–5991.
[40] L. Zuo, M. Jing, J. Li, L. Zhu, K. Lu, Y. Yang, Challenging tough samples in unsupervised domain adaptation, Pattern Recognit. 110 (2020) 107540.

Dayan Guan received his Ph.D. from the Zhejiang University, China. He is currently a Research Fellow with School of Computer Science and Engineering, the Nanyang Technological University, Singapore. His major research interests include computer vision, pattern recognition and machine learning.

Huang Jiaxing received his B.Eng. and M.Sc. in EEE from the University of Glasgow, UK, and the Nanyang Technological University (NTU), Singapore, respectively. He is currently a Research Associate and Ph.D. student with School of Computer Science and Engineering, NTU, Singapore. His research interests include computer vision and machine learning.

Shijian Lu received his Ph.D. from the National University of Singapore. He is currently an Assistant Professor with School of Computer Science and Engineering, the Nanyang Technological University, Singapore. His major research interests include image and video analytics, visual intelligence, and machine learning.

Aoran Xiao received the B.E. and M.E. degrees from Wuhan University, China. He is currently pursuing the Ph.D. degree with School of Computer Science and Engineering, Nanyang Technology University, Singapore. He is also working as research associate in SCALE@NTU. His research interests include computer vision and point cloud data processing.