BiMaL: Bijective Maximum Likelihood Approach to Domain Adaptation in Semantic Scene Segmentation

Thanh-Dat Truong\textsuperscript{1}, Chi Nhan Duong\textsuperscript{2}, Ngan Le\textsuperscript{1}, Son Lam Phung\textsuperscript{4}, Chase Rainwater\textsuperscript{3}, Khoa Luu\textsuperscript{1}
\textsuperscript{1}CVIU Lab, University of Arkansas \quad \textsuperscript{2}Concordia University \quad \textsuperscript{3}Dep. of Industrial Engineering, University of Arkansas \quad \textsuperscript{4}University of Wollongong
\{tt032, thile, cer, khoaluu\}@uark.edu, dcnhan@ieee.org, phung@uow.edu.au

Abstract

Semantic segmentation aims to predict pixel-level labels. It has become a popular task in various computer vision applications. While fully supervised segmentation methods have achieved high accuracy on large-scale vision datasets, they are unable to generalize on a new test environment or a new domain well. In this work, we first introduce a new Unaligned Domain Score to measure the efficiency of a learned model on a new target domain in unsupervised manner. Then, we present the new Bijective Maximum Likelihood (BiMaL) loss that is a generalized form of the Adversarial Entropy Minimization without any assumption about pixel independence. We have evaluated the proposed BiMaL on two domains. The proposed BiMaL approach consistently outperforms the SOTA methods on empirical experiments on “SYNTHIA to Cityscapes”, “GTA5 to Cityscapes”, and “SYNTHIA to Vistas”.

1. Introduction

Semantic segmentation is one of the most popular computer vision topics, which aims to assign each pixel in an image to a predefined class. It has various practical applications, especially in autonomous driving where a segmentation model is needed to recognize roads, sidewalks, pedestrians or vehicles in a large variety of urban conditions. A typical supervised segmentation model is usually trained on datasets with labels. However, annotating images for the semantic segmentation task is costly and time-consuming. Alternatively, a powerful and cost-effective way to acquire a large-scale training set is to use a simulation, e.g. game engines, to create a synthetic dataset \cite{42, 43}. However, fully supervised models \cite{3, 24} trained on the synthetic datasets are often unable to perform well on real images due to the pixel appearance gap between synthetic and real images.

Unsupervised Domain Adaptation (UDA) aims to train a machine learning model on an annotated dataset, i.e. the source, and guarantee its high performance on a new unlabeled dataset, i.e. the target. The UDA approaches have been applied to various computer vision tasks such as Semantic Segmentation \cite{3, 24, 26, 54, 55, 57}, Face Recognition \cite{12, 32, 33, 34, 35}. The recent UDA methods aim to reduce the cross-domain discrepancy, along with the supervised training on the source domain \cite{5, 16, 29, 40, 52, 54}. In particular, these methods aim to minimize the distribution discrepancy of the deep representations extracted from the source and the target domains. This process can be performed at single or multiple levels of deep features using maximum mean discrepancies \cite{16, 29, 52}, or adversarial training \cite{5, 6, 20, 21, 22, 50}. The approaches in this group have shown their potential in aligning the predicted outputs of images from the two domains. However, the binary cross-entropy label predicted by the learned discriminator is usually a weak indication of structural learning for the segmentation task. Another approach named self-training utilizes the pseudo-labels or generative networks conditioned on target images \cite{37, 58}. Semi-supervised learning is an approach related to UDA where the training set consists of both labeled and unlabeled samples. Thus, it has motivated several UDA approaches such as Class-balanced self-training (CBST) \cite{60}, and entropy minimization \cite{4, 17, 40, 47, 54}. Although metrics such as entropy can be efficiently computed and adopted for training, they tend to rely on easy predictions, i.e. high confident scores, as references for the label transfer from source to target domains. This issue is alleviated in a later approach \cite{41} by preventing learned models from over-focusing on high confident areas. However, this type of metrics is formulated in pixel-wised manner, and, therefore, neglects the structural information presented in the image (see Figure 1).

Contributions of this Work. This work presents a new unsupervised domain adaptation approach to tackle the semantic segmentation problem. Table 1 summarizes the difference between our proposed approach and the prior ones. Our contributions can be summarized as follows.

\url{https://github.com/uark-cviu/BiMaL}
Figure 1. Two images have the same entropy but one has a poor prediction (a top image) and one has an better prediction (a bottom image). Columns 1 and 2 are an input image and a ground truth. Columns 3 and 4 are an entropy map and a prediction of AdvEnt [54]. Column 5 is the results of our proposed method. The two predictions produced by AdvEnt have similar entropy scores (0.13 and 0.14). Meanwhile, the BiMaL value of the bottom prediction (0.06) is smaller than the top prediction (0.14). Our results in the last column, which have better BiMaL values than AdvEnt, can well model the structure of an image. In particular, our results have sharper results of a barrier and a rider (white dash box), and a clear boundary between road and sidewalk.

Firstly, a new Unaligned Domain Score (UDS) is introduced to measure the efficiency of the learned model on a target domain in an unsupervised manner. Secondly, the presented UDS is further extended as a new loss function, named Bijective Maximum Likelihood (BiMaL) loss, that can be used with an unsupervised deep neural network to generalize on target domains. Indeed, we further demonstrate BiMaL loss is a generalized form of the Adversarial Entropy Minimization (AdvEnt) [54] without pixel independence assumption. Far apart from AdvEnt that assumes pixel independence, BiMaL loss is formed using a Maximum-likelihood formulation to model the global structure of a segmentation input and a Bijective function to map that segmentation structure to a deep latent space. Finally, the proposed BiMaL method is evaluated on three popular large-scale semantic segmentation benchmarks, including GTA5 [42] to CityScapes [7], SYNTHIA [43] to Cityscapes, and SYNTHIA to Vista [38]. The experimental results demonstrate our proposed BiMaL approach consistently outperforms the State-of-the-Art (SOTA) methods [5, 40, 50, 54, 55] in all these benchmark databases. To the best of our knowledge, this is one of the first works that introduces a novel bijective maximum likelihood approach with flow-based metric to unsupervised domain adaptation in semantic segmentation.

2. Related Work

Unsupervised Domain Adaptation has recently become one of the most active research topics. The common UDA approaches are domain discrepancy minimization [16, 29, 52], adversarial learning [8, 6, 21, 22, 50], entropy minimization [37, 40, 54, 58], self-training [60]. In the scope of this work, UDA is focused on semantic segmentation.

Adversarial training is the most common approach employed to UDA for semantic segmentation. Similar to generative adversarial networks (GANs), the adversarial training paradigm aims at training a discriminator to predict the domain of inputs while the segmentation network tries to fool the discriminator. This adversarial step is trained simultaneously with the supervised segmentation task on the source domain. Hoffman et al. [21] first introduced GAN-based UDA approach to semantic segmentation. Later, Chen et al. [6] presented global and class-wise adaptation learned by adversarial learning on pseudo labels. Considering the difference in spatial distribution, [5] proposed a spatial-aware adaptation method to align two domains along with a target guided distillation loss. Hong et al. [22] learned a conditional generator to transform the feature maps of source domain to be similar to target domain. Tasi et al. [50] used adversarial learning to learn a consistency of scene layout and local context between source and target domains. There are some prior methods that utilize the generative networks to synthesize target images conditioned on source images [58, 37]. Hoffman et al. [20] presented Cycle-Consistent Adversarial Domain Adaptation that aligns at both pixel-level and feature-level representations. Zhu et al. [59] introduced a conservative loss in an adversarial framework that penalizes the easy and hard source examples. We et al. [56] proposed a DCAN framework that uses the channel-wise feature alignment in the segmentation networks. Sakaridis et al. [44] proposed an UDA framework on scene understanding that gradually adapts a segmentation model from non-foggy to heavy-foggy images.

To enhance the performance of domain adaptation, several methods explore the use of privileged information available on source data [2, 27, 45]. Vapnik et al. [53] first introduced the idea of privileged information, i.e. additional information only available at the training process. Later, many methods [19, 30, 36, 46] take advantage of privileged information for various tasks. In semantic segmentation, SPIGAN [25] proposed an UDA approach that utilizes the depth information during the training phase. Following SPI-
GAN, Vu et al. [55] presented an adversarial approach that utilizes the depth-aware source and target images.

Entropy minimization has been used for semi-supervised learning [17, 47]. Vu et al. [54] first introduced the entropy minimization approach for domain adaptation in semantic segmentation. The minimization process is solved by adversarial learning. Later, [40] introduced an intra-domain adaptation approach based on the entropy level of predictions. The learning process involves two phases. The first phase performs adaptation from the source domain to the target domain, whereas the second phase aligns the hard split and easy split within the target domain. Another recent UDA approach is self-training, where the predictions of the trained model are used as pseudo-labels for the unlabeled data. This approach is unsupervised. 

### 3. Unaligned Domain Scores (UDS)

Let \( x_t \in \mathcal{X}_t \subset \mathbb{R}^{H \times W \times 3} \) be an input image of the source domain \((H \text{ and } W \text{ are the height and width of an image})\), and \( x_t \in \mathcal{X}_t \subset \mathbb{R}^{H \times W \times 3} \) be an input image of the target domain, \( G : \mathcal{X} \rightarrow \mathcal{Y} \) where \( \mathcal{X} = \mathcal{X}_s \cup \mathcal{X}_t \) is a semantic segmentation function that maps an input image to its corresponding segmentation map \( y \subseteq \mathbb{R}^{H \times W \times C} \), i.e. \( y = G(x, \theta) \) (\( C \) is the number of semantic classes). In general, given \( N \) labeled training samples from a source domain \( D_s = \{x^i_s, y^i_s\}_1^N \) and \( M \) unlabeled samples from a target domain \( D_t = \{x^i_t\}_1^M \), the unsupervised domain adaptation for semantic segmentation can be formulated as:

\[
\theta^* = \arg\min_{\theta} \sum_{ij} \left[ \mathcal{L}_s(G(x^i_s, \theta), y^i_s) + \mathcal{L}_t(G(x^i_t, \theta)) \right]
\]

where \( \theta \) are the parameters of \( G \), \( p(\cdot) \) is the probability density function. As the labels for \( D_s \) are available, \( \mathcal{L}_s \) can be efficiently formulated as a supervised cross-entropy loss:

\[
\mathcal{L}_s(y_s, \hat{y}_s) = - \sum_{h,w,c} y^h_w_c \log \left( \hat{y}^h_w_c \right)
\]

where \( \hat{y}^h_w_c \) and \( y^h_w_c \) represent the predicted and ground-truth probabilities of the pixel at the location of \((h, w)\) taking the label of \( c \), respectively. Meanwhile, \( \mathcal{L}_t \) handles unlabeled data from the target domain where the ground-truth labels are not available. To alleviate this label lacking issue, several forms of \( \mathcal{L}_t(y_t) \) have been exploited such as cross-entropy loss with pseudo-labels [60]. Probability Distribution Divergence (i.e. Adversarial loss defined via an additional Discriminator) [50, 51], or entropy formulation [54, 40].

### Entropy minimization revisited.

By adopting the Shannon entropy formulation to the target prediction and constraining function \( G \) to produce a high-confident prediction, \( \mathcal{L}_t \) can be formulated as:

\[
\mathcal{L}_t(y_t) = -\frac{1}{\log(C)} \sum_{h,w,c} y^h_w_c \log \left( \hat{y}^h_w_c \right).
\]

Although this form of \( \mathcal{L}_t \) can give a direct assessment of the predicted segmentation maps, it tends to be dominated by the high probability areas (since the high probability areas produce a higher value updated gradient due to \( \frac{\partial \mathcal{L}_t(y_t)}{\partial y^h_w_c} = \frac{1}{\log(C)} \) and \( \lim_{y^h_w_c \to 0} \frac{\partial \mathcal{L}_t(y_t)}{\partial y^h_w_c} = -\infty \), i.e. easy classes, rather than difficult classes [54]. More importantly, this is essentially a pixel-wise formation, where pixels are treated independently of each other. Consequently, the structural information is usually neglected in this form. This issue could lead to a confusion point during training process where two predicted segmentation maps have similar entropy but different segmentation accuracy, one correct and other incorrect as shown in Fig[

### Table 1. Comparison in the properties between our proposed approach and other methods.

| Methods    | Architecture | Source Label | Learning Mechanism | Loss Function | Structural Learning |
|------------|--------------|--------------|--------------------|---------------|---------------------|
| AdaptSeg   | CNN + GAN    | Seg          | Domain Adaptation  | \( \ell_{cde} \) | Weak (binary label)  |
| AdaptPatch | CNN + GAN    | Seg          | Domain Adaptation  | \( \ell_{cde} \) | Weak (binary label)  |
| CBST       | CNN          | Seg          | Self-Training      | \( \ell_{cde} \) | Not Applicable       |
| ADVENT     | CNN + GAN    | Seg          | Domain Adaptation  | EntMin        | Weak (binary label)  |
| MaxSquare  | CNN + GAN    | Seg          | Domain Adaptation  | Squares loss + IW | Weak (binary label)  |
| IntraDA    | CNN + GAN    | Seg          | Curriculum Learning| EntMin        | Weak (binary label)  |
| SPIGAN     | CNN + GAN    | Seg + Depth  | Domain Adaptation  | \( \ell_{cde} + \ell^1 \) | Weak (binary label)  |
| DADA       | CNN + GAN    | Seg + Depth  | Domain Adaptation  | \( \ell_{cde} + \ell_{huber} \) | Depth-aware Label    |
| BiMaL      | CNN + BiN    | Seg          | Domain Adaptation  | Maximum Likelihood | Segmentation Density (Unsupervised) |

Note: \( \ell \) represents the loss function, \( C \) is the number of classes.
y maps due to lighting, scenes, weather), their segmentation significantly in image space (e.g. difference in pixel appearance) segmentation labels of the source domains for q can practically adopt the prior knowledge learned from walks, vehicles should be on roads, etc.). Therefore, one constraints (sky has to be above roads, trees should be on side-class distributions as well as global and local structural con-

Figure 2. The Proposed Framework. The RGB image input is firstly forwarded to a deep semantic segmentation network to produce a segmentation map. The supervised loss is employed on the source training samples. Meanwhile, the predicted segmentation on target training samples will be mapped to the latent space to compute the Bijective Maximum Likelihood loss. The bijective mapping network is trained on the ground-truth images of the source domain.

### 3.1. The Proposed UDS Metric

In the entropy formulation, the pixel independent constraints are employed to convert the image-level metric to pixel-level metric. In contrast, we propose an image-level UDS metric that can directly evaluate the structural quality of y_t. Particularly, let p_t(y_t) and q_t(y_t) be the probability mass functions of the predicted distribution and the real (actual) distribution of the predicted segmentation map y_t, respectively. UDS metric measuring the efficiency of function G on the target dataset can be expressed as follows:

$$\text{UDS} = \mathbb{E}_{y_t \sim p_t(y_t)} \mathcal{L}_Y (p_t(y_t), q_t(y_t))$$

$$= \int \mathcal{L}_Y (p_t(y_t), q_t(y_t)) p_t(y_t) dy_t,$$

where $$\mathcal{L}_Y (p_t(y_t), q_t(y_t))$$ defines the distance between two distributions $$p_t(y_t)$$ and $$q_t(y_t)$$. Since there is no label for sample in the target domain, the direct access to $$q_t(y_t)$$ is not available. Note that although x_s and x_t could vary significantly in image space (e.g. difference in pixel appearance due to lighting, scenes, weather), their segmentation maps y_t and y_s share similar distributions in terms of both class distributions as well as global and local structural constraints (sky has to be above roads, trees should be on sidewalks, vehicles should be on roads, etc.). Therefore, one can practically adopt the prior knowledge learned from segmentation labels of the source domains for $$q_t(y_t)$$ as

$$\text{UDS} \approx \int \mathcal{L}_Y (p_t(y_t), q_s(y_t)) p_t(y_t) dy_t,$$

where the distribution $$q_s(y_t)$$ is the probability mass functions of the real distribution learned from ground-truth segmentation maps of D_s. As a result, the proposed USD metric can be computed without the requirement of labeled target data for learning the density of segmentation maps in target domain. There are several choices for $$\mathcal{L}_Y$$ to estimate the divergence between the two distributions $$p_t(y_t)$$ and $$q_s(y_t)$$. In this paper, we adopt the common metric such as Kullback–Leibler (KL) formula for $$\mathcal{L}_Y$$. Note that other metrics are also applicable in the proposed USD formulation. Moreover, to enhance the smoothness of the predicted semantic segmentation, a regularization term $$\tau$$ is imposed into $$\mathcal{L}_Y$$ as

$$\mathcal{L}_Y (p_t(y_t), q_s(y_t)) = \log \left( \frac{p_t(y_t)}{q_s(y_t)} \right) + \tau(y_t).$$

By computing UDS, one can measure the quality of the predicted segmentation maps y_t on the target data.

In the next sections, we firstly discuss in details the learning process of $$q_s(y_t)$$, and then derivations of the UDS metric for the novel Bijective Maximum Likelihood loss.

### 3.2. Learning Distribution with Bijective Mapping on the Source Domain

Let $$F : Y \rightarrow Z$$ be the bijective mapping function that maps a segmentation $$\hat{y}_s \in Y$$ to the latent space Z, i.e. $$\hat{z}_s = F(\hat{y}_s, \theta_F)$$, where $$\hat{z}_s \sim q_z(\hat{z}_s)$$ is the latent variable, and $$q_z$$ is the prior distribution. Then, the probability distribution $$q_s(\hat{y}_s)$$ can be formulated via the change of variable formula:

$$\log(q_s(\hat{y}_s)) = \log(q_z(\hat{z}_s)) + \log \left( \left| \frac{\partial F(\hat{y}_s, \theta_F)}{\partial y_s} \right| \right),$$

where $$\theta_F$$ is the parameters of F, $$\left| \frac{\partial F(\hat{y}_s, \theta_F)}{\partial y_s} \right|$$ denotes the Jacobian determinant of function $$F(\hat{y}_s, \theta_F)$$ with respect to $$\hat{y}_s$$. To learn the mapping function, the negative log-likelihood will be minimized as follows:

$$\theta_F^* = \arg \min_{\theta_F} \mathbb{E}_{\hat{z}_s \sim q_z(\hat{z}_s)} \left[ - \log(q_s(\hat{y}_s)) \right]$$

$$= \arg \min_{\theta_F} \mathbb{E}_{\hat{z}_s \sim q_z(\hat{z}_s)} \left[ - \log(q_s(\hat{y}_s)) - \log \left( \left| \frac{\partial F(\hat{y}_s, \theta_F)}{\partial y_s} \right| \right) \right].$$
In general, there are various choices for the prior distribution \( q_s \). However, the ideal distribution should satisfy two criteria: (1) simplicity in the density estimation, and (2) easy in sampling. Considering the two criteria, we choose Normal distribution as the prior distribution \( q_s \). Note that any other distribution is also feasible as long as it satisfies the mentioned criteria.

To enforce the information flow from a segmentation domain to a latent space with different abstraction levels, the bijective function \( F \) can be further formulated as a composition of several sub-bijective functions \( f_i \) as \( F = f_1 \circ f_2 \circ \ldots \circ f_K \), where \( K \) is the number of sub-functions. The Jacobian \( \frac{\partial F}{\partial y_i} \), can be derived by \( \frac{\partial F}{\partial y_i} = \frac{\partial f_1}{\partial y_i} \cdot \frac{\partial f_2}{\partial y_1} \ldots \frac{\partial f_K}{\partial y_{K-1}} \).

With this structure, the properties of each \( f_i \) will define the properties for the whole bijective mapping function \( F \). Interestingly, with this form, \( F \) becomes a non-linear function built from a composition of convolutional layers. Several DNN structures \([8, 9, 19, 15, 23, 14, 49]\) can be adopted for sub-functions.

### 4. Bijective Maximum Likelihood Loss

In this section, we present the proposed Bijective Maximum Likelihood (BiMaL) which can be used as the loss of target domain \( \mathcal{L}_t \). From Eqns. (5) and (6), UDS metric can be rewritten as follows:

\[
\text{UDS} = \int \left[ \log \left( \frac{p_t(y_t)}{q_s(y_t)} \right) + \tau(y_t) \right] p_t(y_t)dy_t
\]

\[
= E_{y_t \sim p_t(y_t)} \left[ \log (p_t(y_t)) \right] - E_{y_t \sim p_t(y_t)} \left[ \log (q_s(y_t)) \right] + E_{y_t \sim p_t(y_t)} \left[ \tau(y_t) \right]
\]

\[
\leq E_{y_t \sim p_t(y_t)} \left[ - \log (q_s(y_t)) + \tau(y_t) \right]
\]

It should be noticed that with any form of the distribution \( p_t \), the above inequality still holds as \( p_t(y_t) \in [0, 1] \) and \( \log(p_t(y_t)) \leq 0 \). Now, we define our Bijective Maximum Likelihood Loss as

\[
\mathcal{L}_t(y_t) = - \log(q_s(y_t)) + \tau(y_t),
\]

where \( \log(q_s(y_t)) \) defines the log-likelihood of \( y_t \) with respect to the density function \( q_s \). Then, by adopting the bijective function \( F \) learned from Eqn. (8) using samples from source domain and the prior distribution \( q_s \), the first term of \( \mathcal{L}_t(y_t) \) in Eqn. (10) can be efficiently computed via log-likelihood formulation:

\[
\mathcal{L}_{lik}(y_t) = - \log(q_s(y_t))
\]

\[
= - \log(q_s(z_t)) - \log \left( \left| \frac{\partial F(y_t, \theta_F)}{\partial y_t} \right| \right),
\]

where \( z_t = F(y_t, \theta_F) \). Thanks to the bijective property of the mapping function \( F \), the minimum negative log-likelihood loss \( \mathcal{L}_{lik}(y_t) \) can be effectively computed via the density of the prior distribution \( q_z \) and its associated Jacobian determinant \( \left| \frac{\partial F(y_t, \theta_F)}{\partial y_t} \right| \). For the second term of \( \mathcal{L}_t(y_t) \), we further enhance the smoothness of the predicted semantic segmentation with the pair-wised formulation to encourage similar predictions for neighbourhood pixels with similar color:

\[
\tau(y_t) = \sum_{h,w,h',w'} \exp \left\{ - \frac{|y_t^{h,w} - y_t^{h',w'}|^2}{2\sigma_1^2} - \frac{|y_t^{h,w} - y_t^{h',w'}|^2}{2\sigma_2^2} \right\}
\]

where \( (h', w') \in N_{h,w} \) denotes the neighborhood pixels of \( (h, w) \), \( x^{h,w} \) represents the color at pixel \( (h, w) \); and \( \{\sigma_1, \sigma_2\} \) are the hyper parameters controlling the scale of Gaussian kernels. It should be noted that any regularizers \([3, 13]\) enhancing the smoothness of the segmentation results can also be adopted for \( \tau \). Putting Eqns. (10), (11) to Eqn (1), the objective function can be rewritten as:

\[
\theta^* = \arg \min_{\theta} \mathbb{E}_{y_t \sim p(y_t), \hat{y}_s \sim p(\hat{y}_s)} \left[ \mathcal{L}_t(y_t, \hat{y}_s) \right] + \mathbb{E}_{y_t \sim p(y_t)} \left[ \mathcal{L}_{lik}(y_t + \tau(y_t)) \right]
\]

Figure 2 illustrates our proposed BiMaL framework to learn the deep segmentation network \( G \). Also, we can prove that direct entropy minimization as Eqn. (3) is just a particular case of our log likelihood maximization. We will further discuss how our maximum likelihood can cover the case of pixel-independent entropy minimization in Section 4.2.

### 4.1. BiMaL properties

**Global Structure Learning.** Sharing similar property with \([10, 11, 15, 39, 48]\), from Eqn. (7), as the learned density function is adopted for the entire segmentation map \( \hat{y}_s \), the global structure in \( \hat{y}_s \) can be efficiently captured and modeled.

**Tractability and Invertibility.** Thanks to the designed bijection \( F \), the complex distribution of segmentation maps can be efficiently captured. Moreover, the mapping function is bijective, and, therefore, both inference and generation are exact and tractable.

### 4.2. Relation to Entropy Minimization

The first term of UDS in Eqn. (9) can be derived as

\[
\int \log \left( \frac{p_t(y)}{q_s(y)} \right) p_t(y)dy_t \geq 0
\]

\[
\equiv \mathbb{E}_{y_t \sim p_t(y_t)} \left[ \log (p_t(y_t)) - \log (q_s(y_t)) \right] \geq 0
\]

\[
\equiv \mathbb{E}_{y_t \sim p_t(y_t)} \left[ - \log (q_s(y_t)) \right] \mathbb{E}_{y_t \sim p_t(y_t)} \left[ - \log (p_t(y_t)) \right]
\]

\[
\equiv \mathbb{E}_{y_t \sim p_t(y_t)} \left[ \mathcal{L}_{lik}(y_t) \right] \geq \text{Ent}(Y_t)
\]

where \( Y_t \) is the random variable with possible values \( y_t \sim p_t(y_t) \), and \( \text{Ent}(Y_t) \) denotes the entropy of the random variable \( Y_t \). It can be seen that the proposed negative log-likelihood \( \mathcal{L}_{lik} \) is an upper bound of the entropy of \( Y_t \). Therefore, minimizing our proposed BiMaL loss will also
enforce the entropy minimization process. Moreover, by not assuming pixel independence, our proposed BiMaL can model and evaluate structural information at the image-level better than previous pixel-level approaches [4, 40, 54].

5. Experimental Results

This section will present our experimental results on three different benchmarks, i.e. SYNTHIA to Cityscapes, GTA to Cityscapes, and SYNTHIA to Vistas. First, we overview datasets and network architectures used in our experiments. Second, we present the ablation study to analyze the effectiveness of our proposed BiMaL and the capability of the bijective network. Finally, we present the quantitative and qualitative results of our method compared to prior methods on the three benchmarks.

5.1. Datasets

**GTA5** [42] is a synthetic dataset containing 24,966 densely labelled images at the resolution of 1,914 × 1,052. This dataset was collected from the game Grand Theft Auto V. The ground-truth annotations were automatically generated with 33 categories. In our experiments, we consider 19 categories that are compatible with the Cityscapes [7].

**SYNTHIA (SYNHIA-RAND-CITYSCAPES) [43]** is also synthetic dataset that contains 9,400 pixel-level labelled RGB images. In our experiments, we use the 16 common categories that overlap with the Cityscapes dataset.

**Cityscapes** [7] is a real-world dataset including 3,975 images with fine semantic, dense pixel annotations of 30 classes. In our experiments, 2,495 images are used for training and 500 images are used for testing.

**Vistas (Mapillary Vistas Dataset) [38]** is diverse street-level imagery dataset with pixel-accurate and instance-specific human annotations for understanding street scenes around the world. Vistas consists of 25,000 high-resolution images and 124 semantic object categories.

In our experiments, we consider 7 classes that are common to SYNTHIA, Cityscapes and Vistas as shown in Fig. 3.

5.2. Network Architectures

In our experiments, we adopt the DeepLab-V2 [3] with ResNet-101 [18] backbone for the segmentation network $G$. Also, we utilize the Atrous Spatial Pyramid Pooling with sampling rate $\{6, 12, 18, 24\}$. We only use the output of layer $\text{conv}5$ to predict the segmentation. In the Bijective network $F$, we use the multi-scale architecture as [8, 9, 14, 23, 39]. For each scale, we have multiple steps of flow, each of which consists of ActNorm, Invertible $1 \times 1$ Convolution, and Affine Coupling Layer [23, 38]. In our experiments, the number of scales and number of flow steps are set to 4 and 32, respectively.

The entire framework is implemented in PyTorch [41]. Training and validating models are conducted on 4 GPUs of NVIDIA Quadpro P8000 with 48GB each GPU. Segmentation and bijective networks are trained by a Stochastic Gradient Descent optimizer [1] with learning rate $2.5 \times 10^{-4}$.
momentum 0.9, and weight decay $10^{-4}$. The batch size per GPU is set to 4 for segmentation network, and 16 for learning bijective network. The image size is set to $1280 \times 720$ pixels in all experiments.

### 5.3. Ablation Study

**Effectiveness of Losses.** Figure 4 reports the semantic performance (mIoU) of BiMaL on the 16 classes of the Cityscapes validation set when the model is trained on SYNTHIA dataset. We consider three cases: (1) without adaptation (train with source only), (2) BiMaL without regularization term ($\mathcal{L}_{lik}(y)$ only), and (3) BiMaL with regularization term ($\mathcal{L}_{lik}(y) + \tau(y)$). Overall, the proposed BiMaL improves the performance of the method. In particular, the mIoU accuracy of the baseline (without adaptation) is 33.7%. In comparison, BiMaL, without regularization and BiMaL with regularization achieve the mIoU accuracy of 43.5% and 46.2%, respectively. In terms of per-class accuracy, using BiMaL significantly improves the performance on classes of ‘road’, ‘sidewalk’, ‘bus’, and ‘motorcycle’.

**Bijective Network Ability.** We conduct a pilot experiment of the bijective network on ground-truth semantic segmentation images of the GTA dataset. This experiment aims to analyze the ability of the bijective network in modeling the image and structure information. The number of scales and number of flow steps are set to 3, and 32, respectively. As shown in Figure 5(a), our bijective network can successfully reconstruct good-quality images. It also can synthesize images sampled from the latent space as shown in Figure 5(b). These experimental results have shown that the bijective network can model images even with complex structures as scene segmentation.

### 5.4. Comparisons with SOTA Methods

We present the experimental results of the proposed approach in comparison to other strong baselines. Comparative experiments are conducted on three benchmarks: i.e. SYNTHIA to Cityscapes, GTA5 to Cityscapes, and SYNTHIA to Vistas. In all three benchmarks, our method consistently achieves the SOTA semantic segmentation performance in term of “mean Intersection over Union” (mIoU).

**SYNTHIA to Cityscapes.** Table 2 presents the semantic performance (mIoU) on the 16 classes of the Cityscapes validation set of different models trained on GTA5.

**GTA5 to Cityscapes.** Table 3 shows the mIoU of 19 classes of Cityscapes on the validation set. Our proposed method achieves better accuracy than the prior methods, i.e. 46.2% higher than DADA by 3.6%. Considering per-class results, our method significantly improves the results on classes of ‘sidewalk’ (51.5%), ‘car’ (85.7%), and ‘bus’ (44.5%). We also report the results on a 13-class subset where our proposed method also achieves the State-of-the-Art performance.

**GTA5 to Cityscapes.** Table 5 shows the mIoU of 19 classes of Cityscapes on the validation set. Our approach gains mIoU of 47.3% that is state-of-the-art performance compared to the prior methods. Analysing per-class results, our method gains the improvement on most classes. In particular, the results on classes of ‘terrain’ (+10.0%), ‘truck’ (+9.1%), ‘bus’ (+8.0%), ‘motorbike’ (+5.6%) demonstrate significant improvements compared to AdvEnt. For other classes, the proposed method gains moderate improvements, compared to prior SOTA methods.
SYNTHIA to Vistas. Table 4 reports the mIoU on 7 classes of the Vistas testing set. Our approach gains an mIoU of 67.2% which is the SOTA performance compared to prior methods. Moreover, our method also gains moderate improvements in per-class accuracy.

Qualitative Results. Figure 6 illustrates the qualitative results of the SYNTHIA to Cityscapes experiment. Our method gives the better qualitative results compared to a model trained on the source domain and AdvEnt [54]. Our method can model well the structure of an image. In particular, our results have a clear border between ‘road’ and ‘sidewalk’. Meanwhile, the results of model trained on source only and AdvEnt have an unclear border between ‘road’ and ‘sidewalk’. Overall, our qualitative semantic segmentation results are sharper than the results of AdvEnt.

6. Conclusions

This paper has presented a new Bijective Maximum Likelihood approach to domain adaptation in semantic scene segmentation. Compared to Adversarial Entropy Minimization loss, it is a more generalized form and can work without any assumption about pixel independence. In addition, a new Unaligned Domain Score metric has been also introduced to measure the efficiency of a segmentation model on a new target domain in the unsupervised manner. Through intensive experiments on three different datasets, i.e. SYNTHIA to Cityscapes, GTA to Cityscapes, and SYNTHIA to Vistas, we achieve SOTA performance compared to prior methods. Specifically, our semantic segmentation accuracy on these three benchmarks are 46.2%, 47.3%, and 68.4%, respectively. The future direction of this work is to solve challenging cases coming from the differences in “segmentation structures” between source and target domains such as left- and right-hand traffic.

Acknowledgement This work is supported by NSF Data Science, Data Analytics that are Robust and Trusted (DART), Chancellor’s Innovation Fund, UAF and NSF Small Business Grant.
References

[1] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*. 2010.
[2] Lin Chen, Wen Li, and Dong Xu. Recognizing RGB images by learning from RGB-D data. In *CVPR*, 2014.
[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *TPAMI*, 2018.
[4] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *ICCV*, 2019.
[5] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *CVPR*, 2018.
[6] 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 *ICCV*, 2017.
[7] 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 *CVPR*, 2016.
[8] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation, 2015.
[9] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp, 2017.
[10] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Tien D. Bui. Longitudinal face modeling via temporal deep restricted boltzmann machines. In *CVPR*, 2016.
[11] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Tien D. Bui. Deep appearance models: A deep boltzmann machine approach for face modeling. *IJCV*, 2019.
[12] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Ngan Le. Shrinkteamet: Million-scale lightweight face recognition via shrinking teacher-student networks. *arXiv:1905.10620*, 2019.
[13] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, Nghia Nguyen, Eric Patterson, Tien D. Bui, and Ngan Le. Automatic face aging in videos via deep reinforcement learning. In *CVPR*, 2019.
[14] Chi Nhan Duong, Kha Gia Quach, Khoa Luu, T Hoang Ngan Le, Marios Savvides, and Tien D Bui. Learning from longitudinal face demonstration—where tractable deep modeling meets inverse reinforcement learning. *IJCV*, 2019.
[15] Chi Nhan Duong, Thanh-Dat Truong, Khoa Luu, Kha Gia Quach, Hung Bui, and Kaushik Roy. Vec2face: Unveil human faces from their blackbox features in face recognition. In *CVPR*, 2020.
[16] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.
[17] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *NIPS*, 2005.
[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
[19] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *CVPR*, 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 *ICML*, 2018.
[21] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. FCNs in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv:1612.02649*, 2016.
[22] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *CVPR*, 2018.
[23] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *NIPS*, 2018.
[24] T. Hoang Ngan Le, Kha Gia Quach, Khoa Luu, Chi Nhan Duong, and Marios Savvides. Reformulating level sets as deep recurrent neural network approach to semantic segmentation. *TIP*, 2018.
[25] Kuan-Hui Lee, German Ros, Jie Li, and Adrien Gaidon. SPGAN: Privileged adversarial learning from simulation. In *ICLR*, 2019.
[26] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. Content-consistent matching for domain adaptive semantic segmentation. In *ECCV*, 2020.
[27] Wen Li, Li Niu, and Dong Xu. Exploiting privileged information from web data for image categorization. In *ECCV*, 2014.
[28] Xinzhe Li, Qianru Sun, Yaoao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. Learning to self-train for semi-supervised few-shot classification. In *NeurIPS*, 2019.
[29] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015.
[30] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *ICLR*, 2016.
[31] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. *arXiv:1809.09478*, 2019.
[32] K. Luu, T. D. Bui, K. Ricanek Jr., and C. Y. Suen. Age estimation using active appearance models and support vector machine regression. In *BTAS*, 2009.
[33] K. Luu, T. D. Bui, and C. Y. Suen. Kernel spectral regression of perceived age from hybrid facial features. In *FG*, 2011.
[34] K. Luu, K. Ricanek Jr., T. D. Bui, and C. Y. Suen. The familial face database: A longitudinal study of family-based growth and development on face recognition. In *ROBUST*, 2008.
[35] K. Luu, K. Seshadri, M. Savvides, T. D. Bui, and C. Y. Suen. Contourlet appearance model for facial age estimation. In *IJCB*, 2011.
[36] Taylor Mordan, Nicolas Thome, Gilles Henaff, and Matthieu Cord. Revisiting multi-task learning with rock: a deep residual auxiliary block for visual detection. In NIPS, 2018.

[37] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In CVPR, 2018.

[38] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In ICCV, 2017.

[39] Chi Nhan Duong, Kha gia Quach, Khoa Luu, Ngan Le, and Marios Savvides. Temporal non-volume preserving approach to facial age-progression and age-invariant face recognition. In ICCV, Oct 2017.

[40] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In CVPR, 2020.

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

[42] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In ECCV, 2016.

[43] 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 CVPR, 2016.

[44] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In ECCV, 2018.

[45] Nikolaos Sarafianos, Michalis Vrigkas, and Ioannis A Kakadiaris. Adaptive SVM+: Learning with privileged information for domain adaptation. In ICCV, 2017.

[46] Viktoriaia Sharmanska, Novi Quadrianto, and Christoph H. Lampert. Learning to rank using privileged information. In ICCV, 2013.

[47] Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. ICLR, 2016.

[48] Thanh-Dat Tran, Chi Nhan Duong, Khoa Luu, Minh-Triet Tran, and Ngan Le. Domain generalization via universal non-volume preserving approach. In CRV, 2020.

[49] Thanh-Dat Tran, Chi Nhan Duong, Minh-Triet Tran, Ngan Le, and Khoa Luu. Fast flow reconstruction via robust invertible n × n convolution. Future Internet, 2021.

[50] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulte, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In CVPR, 2018.

[51] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulte, and Manmohan Chandraker. Domain adaptation for structured output via discriminative representations. arXiv:1901.05427, 2019.

[52] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In CVPR, 2017.

[53] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. Neural networks, 2009.

[54] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In CVPR, 2019.

[55] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In ICCV, 2019.

[56] 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 ECCV, 2018.

[57] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. In NIPS, 2019.

[58] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In ICCV, 2017.

[59] Xinge Zhu, Hui Zhou, Ceyuan Yang, Jianping Shi, and Dahua Lin. Penalizing top performers: Conservative loss for semantic segmentation adaptation. In ECCV, 2018.

[60] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In ECCV, 2018.