Generalize then Adapt: Source-Free Domain Adaptive Semantic Segmentation

Jogendra Nath Kundu[1*] Akshay Kulkarni[1*] Amit Singh[1] Varun Jampani[2] R. Venkatesh Babu[1]
[1]Indian Institute of Science, Bangalore [2]Google Research

Abstract

Unsupervised domain adaptation (DA) has gained substantial interest in semantic segmentation. However, almost all prior arts assume concurrent access to both labeled source and unlabeled target, making them unsuitable for scenarios demanding source-free adaptation. In this work[1], we enable source-free DA by partitioning the task into two: a) source-only domain generalization and b) source-free target adaptation. Towards the former, we provide theoretical insights to develop a multi-head framework trained with a virtually extended multi-source dataset, aiming to balance generalization and specificity. Towards the latter, we utilize the multi-head framework to extract reliable target pseudo-labels for self-training. Additionally, we introduce a novel conditional prior-enforcing auto-encoder that discourages spatial irregularities, thereby enhancing the pseudo-label quality. Experiments on the standard GTA5→Cityscapes and SYNTHIA→Cityscapes benchmarks show our superiority even against the non-source-free prior-arts. Further, we show our compatibility with online adaptation enabling deployment in a sequentially changing environment.

1. Introduction

Almost all supervised learning systems assume that the training and testing data follow the same input distribution. However, this assumption is impractical as target scenarios often exhibit a distribution shift. For example, self-driving cars often fail to generalize when deployed in conditions different from training, such as cross-city [10] or cross-weather [59] deployment. This is because the model fails to apprehend the generic, causal factors of variations and instead, holds on to domain-specific spurious correlations [24]. Over-reliance on training data from a particular distribution can cause the model to fail even for mild domain-shifts like changes in illumination, texture, background, *etc*.

Unsupervised domain adaptation (DA) is one of the primary ways to address such problems. Here, the goal is to transfer the knowledge from a labeled source domain to an unlabeled target domain. The major limitation of typical DA approaches [58] is the requirement of concurrent access to both source and target domain samples. While concurrent access better characterizes the distribution shift, it is a major bottleneck for real-world deployment scenarios. Consider a modern corporate dealing where the *vendor* organization has access to a large-scale labeled dataset (*i.e. source-data*) which is used to train a *source-model*. The *vendor* finds multiple *clients* interested in deploying the *source-model* in their specific target environments. However, both parties are restrained from data sharing due to proprietary, privacy, or profit related concerns. This motivates us to seek learning frameworks where the *vendor* can trade only the *source-model* and the client can perform target adaptation without the *source-data*. This special case of domain adaptation [40, 34, 43] is *Source-Free Domain Adaptation* (SFDA).

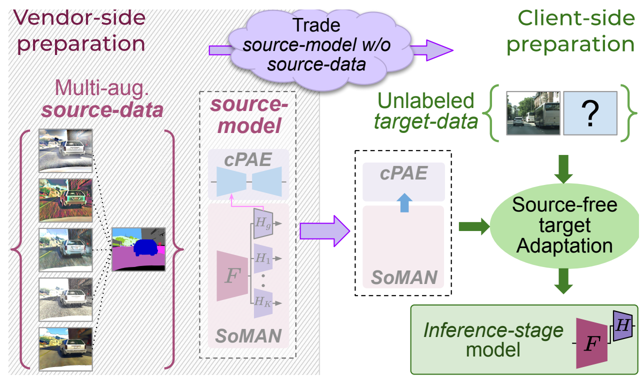

Figure 1. In source-free DA, the vendor accesses source-data to prepare a foresighted source-model. Following this, the client receives only the source-model to perform unsupervised target adaptation while prevented access to the proprietary source-data.

In this work, we aim to develop an SFDA framework for semantic segmentation of urban road scenes. In a co-operative setup, both vendor and the client must adopt specialized learning strategies to benefit the end goal.

a) Vendor-side strategies. These strategies can be discussed under two broad aspects viz. source dataset and training strategy. The vendor must acquire a substantially diverse large-scale dataset aiming to subsume unknown target scenarios. In literature, Multi-Source DA (MSDA) [89, 71, 1] and domain generalization (DG) [38] works use multiple labeled source domains to improve target gener-

*Equal contribution.
[1]Project page: https://sites.google.com/view/sfdaseg

alization. However, gathering annotation for more than one domain is costly and time-consuming [12]. Thus, we focus on developing a strategy to simulate multiple novel domains from samples from a single labeled domain. Carefully crafted augmentations randomly perturb the task-irrelevant factors (such as stylization, texture modulation, *etc*.), facilitating the learning of domain-invariant representations. Hence, we devise multiple augmentation-groups (`AG`s), where each group modulates the image by varying certain statistics thereby constructing virtual, labeled source-domains, to be treated as a multi-source dataset.

Next, we focus on developing an effective training strategy. The naive solution would be to train a single model on the entire multi-source dataset to learn domain-generic features. However, this can lead to sub-optimal performance if a certain `AG` alters the task-relevant causal factors [24]. Further, the target domain may be similar to one or a combination of `AG`s. In such cases, domain-specific (`AG`-specific) learning is more helpful. This motivates us to seek a domain-specific framework to complement the domain-generic model. Thus, we give theoretical insights to analyze domain-specific hypotheses and propose *Source-only Multi-Augmentation Network* (`SoMAN`) as shown in Fig. 1.

Going forward, we recognize that `SoMAN` may lack the ability to capture inductive bias, which would prevent the model from manifesting structurally consistent predictions. This is particularly important for dense prediction tasks [32, 31]. Modeling general context dependent priors encourages the prediction of plausible scene segments while discouraging common irregularities (*e.g*. merged-region or split-region issues [33]). To this end, we introduce a separate model namely, *conditional Prior-enforcing Auto-Encoder* (`cPAE`). `cPAE` is trained on segmentation maps available with the vendor, and used at the client-side to improve the source-free adaptation performance.

b) Client-side strategies. We draw motivation from pseudo-label based self-training approaches [17, 95]. The target samples are passed through the *source-model* to select a set of pseudo-labels which are later used to finetune the network. In the absence of *source-data*, effectiveness of such self-training depends on the following two aspects. First, the training must be regularized to retain the *vendor-side*, *task-specific knowledge*. We address this by allowing only a handful of weights to be updated while others are kept frozen from the *vendor-side* training. Second, the pseudo-label selection criteria must overcome issues related to label-noise and information redundancy. We address this by selecting the optimal prediction from the `SoMAN`-heads and using the pruned output after forwarding through `cPAE`.

In summary, we make the following main contributions:

- We propose to address *source-free* DA by casting the *vendor-side* training as multi-source learning. To this end, we provide theoretical insights to analyze different ways to aggregate the domain-specific hypotheses. It turns out that a combination of domain-generic and *leave-one-out* configuration performs the best.
- While accessing a single source domain, we propose a systematic way to select a minimal set of effective augmentations to resemble a multi-source scenario. The vendor uses this to develop a multi-head network, `SoMAN` subscribing to the leave-one-out configuration.
- Aiming to have a strong support for the spatially-structured segmentation task, we develop a conditional prior-enforcing auto-encoder. This encourages plausible dense predictions thereby enhancing the quality of pseudo-labels to aid the *client-side* self-training.
- Our *source-free* framework achieves *state-of-the-art* results for both GTA5 $\rightarrow$ Cityscapes and SYNTHIA $\rightarrow$ Cityscapes adaptation benchmarks, even when compared against the *non-source-free* prior arts.

2. Related Work

Here, we briefly review the segmentation DA literature [65].

Feature-space DA. The early works in DA for semantic segmentation are inspired from the GAN framework [16], involving training a segmentation network to confuse a domain discriminator enforcing domain invariance on the latent features [20]. Several works [8, 41, 21, 19] utilized this discriminative alignment [93, 9, 46, 14] while adding complementary modules [74, 13, 76] to improve adaptation. Another line of works [60, 7, 48, 66, 79, 81] use the same framework on low-dimensional output space [84, 30, 80, 69, 72, 73] instead of high-dimensional feature space. However, these works require cumbersome adversarial training and rely on source-target co-existence.

Image-space DA. The success of CycleGAN [92] for image-to-image translation led to several DA approaches [41, 19, 9, 15, 52] utilizing it for input-level adaptation while also addressing semantic consistency in the transformed images. Another category of works [3, 57, 87, 11, 78] explore style-transfer techniques for input-level perceptual invariance [77, 83, 47, 37, 75] between source and target domains. However, these works also assume the co-existence of source and target domains.

Source-free DA. Bateson *et al*. [2] perform *source-free* DA for medical segmentation using entropy minimization and class-ratio alignment. Concurrent *source-free* works use data-free distillation, self-training, patch-level self-supervision [44] and feature corruption with entropy regularization [63] focused on target adaptation. In contrast, we develop a novel approach for vendor-side source training.

DA via self-training. Early works [95, 41, 94] use highly confident target predictions as pseudo-labels, selected using a confidence threshold. To improve the pseudo-labels, prior works used prediction ensembling [5, 83, 91, 90], extra networks [11], applied constraints [64], modified the confidence thresholding technique [50, 39, 61], utilized image-level pseudo-labels [55] and intra-domain (easy-hard) adversarial training [53]. Most prior arts use labeled source with self-training to retain *task-specific* source knowledge.

DG and MSDA. [89] use multiple synthetic datasets for Multi-Source DA (MSDA) in segmentation. Restricted to a single source setting, we use data augmentation techniques to generate new domains. In the presence of multi-source data, the vendor-side training is equivalent to domain generalization [85, 6, 54] as it does not involve training on target.

3. Approach

Consider a set of source image and segmentation pairs $(x_s, y_s) \in \mathcal{D}_s$ where the source images x_s are drawn from a marginal distribution p_s. The unlabeled target images $x_t \in \mathcal{D}_t$ are drawn from p_t. However, the output segmentation maps follow a single marginal distribution p_y. The goal is to learn a mapping $\hat{y}_s = h(x_s)$ that can generalize well for x_t. The proposed *source-free* domain adaptation is broadly divided into two: *vendor-side* and *client-side*.

3.1. Vendor-side Strategy

In the absence of target data, the vendor's task effectively reduces to domain generalization (DG) [38]. DG is shown to be highly effective in the presence of multiple source domains. Thus, we plan to cast the vendor-side model preparation as a multi-source representation learning problem.

Non-source-free paradigm. We assume access to K source datasets $(x_{s_i}, y_{s_i}) \in \mathcal{D}_{s_i}\ \forall\ i \in [K] = \{1, 2, \dots, K\}$ where images x_{s_i} are drawn from marginal distribution p_{s_i}. In non-source-free paradigm, the objective is to utilize all the domains (including the target) to realize a hypothesis $h^* = \arg\min_{h \in \mathcal{A}} \epsilon_t(h)$ with a small target error, where

$$\epsilon_t(h) = \underset{(x,y)\sim p_t}{\mathbb{E}} [\mathcal{L}(h(x), y)] \;\; \text{where} \;\; h \in \mathcal{H}^{\alpha^*} \subset \mathcal{A} \quad (1)$$

Here, $\mathcal{L}$ is the loss and $\mathcal{A}$ is the hypothesis space. $\mathcal{H}^{\alpha^*} \subset \mathcal{A}$ can be interpreted as a hypothesis subspace spanning the hypotheses that can be learned using the best convex combination of sources $\alpha^* \in \Delta = \{\alpha \in [0,1]^K : \sum_{i=1}^{K} \alpha[i] = 1\}$ in the presence of concurrent access to $\{\mathcal{D}_{s_i}\}_{i=1}^{K}$ and $\mathcal{D}_t$, *i.e.* $\alpha^* = \arg\min_\alpha(\arg\min_{h \in \mathcal{A}^\alpha} \epsilon_t(h))$.

While operating in a source-free paradigm [35, 36], let the vendor be approached by M number of clients, each with different target domains $t_j\ \forall\ j \in [M]$. For every target t_j, there exists a specific α_j^* such that $\epsilon_{t_j}(h \in \mathcal{H}^{\alpha_j^*}) \leq \epsilon_{t_j}(h \in \mathcal{H}^{\alpha})\ \forall\ \alpha \in \Delta$. However, in the absence of concurrent access to source and target domains (SFDA), it is not possible to optimize for α_j^* for any target t_j. Thus, we propose a source-free multi-domain paradigm.

Definition 1. (Source-free multi-domain paradigm) *Consider a vendor who has access to labeled data $\{\mathcal{D}_{s_i}\}_{i=1}^{K}$ from K source domains and a client who has access to unlabeled target data $\mathcal{D}_{t_j}$. In the source-free paradigm, the vendor prepares a prescient model with an immutable hypothesis support set $\mathcal{A}^{SF}$ (a union of certain hypothesis supports) without any information about t_j. This model is traded with the client for target adaptation without any data sharing.*

In the hypothetical scenario of source-target concurrent access, the client can determine the best α_j^* such that $\epsilon_{t_j}(h \in \mathcal{H}^{\alpha_j^*}) \leq \epsilon_{t_j}(h \in \mathcal{A}^{SF})$. The proposed paradigm not only enables adaptation without any data sharing, but also enables the vendor to prepare a single *source-model* for all future clients. Thus, the process becomes more efficient for both vendor and client in terms of compute and storage.

3.1.1 Multi-source representation learning

Under *source-free*, the vendor's objective would be to realize a learning setup that would generalize to a wide range of unseen targets. While aiming to learn a single hypothesis, empirical risk minimization (ERM) [70] would be the best solution (all domains weighted equally). Consider a scenario where p_{t_j}, *i.e.* marginal distribution of the target t_j, matches with the marginal of one of the source domains. Here, the domain-specific expert for that source domain would definitely outperform the ERM baseline. To this end, a hypothesis support set $\mathcal{A}^{SF}$, *i.e.* a union of certain hypothesis supports, would provide better flexibility for SFDA. With this intent, we discuss the following configurations.

a) ERM. Under ERM configuration, we set $\mathcal{A}^{SF} = \mathcal{H}^{\text{ERM}}$ where $\mathcal{H}^{\text{ERM}}$ is formed with equal weightage to all the multi-source domains *i.e.* $\alpha[i] = \frac{1}{K}\ \forall\, i \in [K]$.

b) Domain-experts++ (DE++). This configuration encompasses a set of $K+1$ hypothesis supports. This includes K number of domain-specific experts alongside one ERM support. Thus, we set $\mathcal{A}^{SF}$ as $\mathcal{A}^{\text{DE++}} = \cup_{i=1}^{K} \mathcal{H}_i^{\text{DE}} \cup \mathcal{H}^{\text{ERM}}$. For i^{th} support $\mathcal{H}_i^{\text{DE}}$, $\alpha_i[i'] = \mathbb{1}_{i'=i}\ \forall\, i' \in [K]$ where $\mathbb{1}$ is the indicator function (1 if input condition is true, else 0).

c) Leave-one-out++ (LO++). It may happen that using a particular source may cause information loss that hinders optimal adaptation for a future target. To improve support for such targets, we introduce leave-one-out (LO) hypothesis support where i^{th} subspace $\mathcal{H}_i^{\text{LO}}$ is formed by leaving one domain out, *i.e.* with $\alpha_i[i'] = \frac{1}{K-1}\mathbb{1}_{i \neq i'}\ \forall\, i' \in [K]$. Similar to *DE++*, *LO++* also includes $K+1$ hypothesis supports, *i.e.* K number of LO supports with one ERM. Thus, we set $\mathcal{A}^{SF}$ as $\mathcal{A}^{\text{LO++}} = \cup_{i=1}^{K} \mathcal{H}_i^{\text{LO}} \cup \mathcal{H}^{\text{ERM}}$.

We include the ERM support, *i.e.* $\mathcal{H}^{\text{ERM}}$, in both *LO++* and *DE++* to provide complementary domain-generic information alongside the different forms of domain-specific information. Here, the individual hypothesis supports are implemented as separate classifier heads trained on a common feature extractor (Sec 3.1.3). Note that we only con-

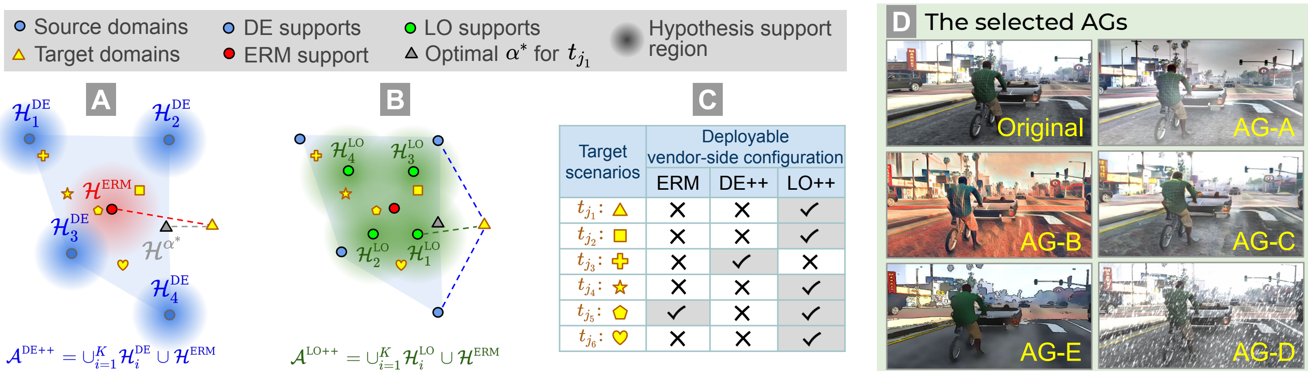

Figure 2. An illustration of hypothesis subspace constituents of **A.** $\mathcal{A}^{\text{DE++}}$ and **B.** $\mathcal{A}^{\text{LO++}}$. The positions of yellow signs represent the best non-source-free hypothesis h_j^* for different target domains t_j. In the source-free paradigm, for each target t_j, the closest vendor hypothesis constituent would provide the support for a reasonable adaptation. **C.** The tick and cross marks for different target scenarios (the rows) denote suitability of the corresponding vendor-side configurations (the columns). For example, $\mathcal{H}^{\text{ERM}}$ is the best support for t_{j_5}, $\mathcal{H}_1^{\text{DE}}$ is the best support for t_{j_3}, $\mathcal{H}_3^{\text{LO}}$ is the best support for t_{j_2}, *etc*. Note that, *LO++* is equipped to reasonably support a wide range of target scenarios. **D.** Visual illustration of selected AGs (augmented domains) based on the proposed augmentation selection criteria.

sider options that require K heads while other domain-specific solutions like leave-r-out have higher computational cost requiring $\binom{K}{r}$ heads. Next, we discuss a result comparing the target error $\epsilon_t(h)$ of the three configurations.

Result 1. *Consider DE++ hypothesis space $\mathcal{A}^{\text{DE++}}$, LO++ hypothesis space $\mathcal{A}^{\text{LO++}}$, and unseen target data $\mathcal{D}_t$. Then,*

$$\begin{aligned} \epsilon_t(h \in \mathcal{A}^{\text{LO++}}) &\leq \epsilon_t(h \in \mathcal{H}^{\text{ERM}}) \\ \epsilon_t(h \in \mathcal{A}^{\text{DE++}}) &\leq \epsilon_t(h \in \mathcal{H}^{\text{ERM}}) \end{aligned} \tag{2}$$

As depicted in Fig. 2, the distributed subspace constituents of $\mathcal{A}^{\text{DE++}}$ and $\mathcal{A}^{\text{LO++}}$ provides better support for a wide range of unknown target domains as compared to the same by $\mathcal{H}^{\text{ERM}}$. Thus, in Eq. 2, $\epsilon_t(h \in \mathcal{H}^{\text{ERM}})$ acts as an upper bound for the target risk, particularly in source-free paradigm. Also, the equality holds as both $\mathcal{A}^{\text{DE++}}$ and $\mathcal{A}^{\text{LO++}}$ already include $\mathcal{H}^{\text{ERM}}$ as a constituent subspace.

Comparison between DE++ and LO++. Though, both *DE++* and *LO++* are better alternatives over ERM, it is not possible to write a general inequality involving only the target errors for *DE++* and *LO++* configurations. Note that, as shown in Fig. 2, for certain target scenarios, target error for DE++ would be less than the same for LO++ and vice versa. However, considering a reasonable domain-shift among the source domains, *LO++* provides lower target error over *DE++* for a wide range of practical target scenarios (see Fig. 2C). *DE++* wins particularly for cases when $p_t \approx p_{s_{i'}}$ for $i' \in [K]$ which is generally quite rare. *LO++* wins for a wide range of unique target scenarios.

3.1.2 Preparing virtual multi-source domains

Having identified *LO++* as the best option, we focus on obtaining the multi-source data. Though, we intend to expand our source-data horizon, we are restricted to a single labeled source domain. Thus, we plan to use diverse data augmentations to simulate a multi-source scenario.

Characterizing multi-domain data. Consider a hypothetical data generation process [56] for the source domain: A data generator ϕ uses the causal class factor f_y and the non-causal domain-related factor f_s to construct a data sample $x_s = \phi(f_y, f_s)$. Next, a set of domain-varying class-preserving augmentations $\{\mathcal{T}_i\}_{i=1}^K$ are applied to obtain,

$$x_{s_i} = \mathcal{T}_i(x_s) = \phi(f_y, f_i + \gamma_i f_s); \quad \gamma_i \in \mathbb{R} \tag{3}$$

Here, $\mathcal{T}_i$ modifies the original domain-specific factor f_s by a weight γ_i (without altering f_y) and introduces a new augmentation related domain-specific factor f_i. Thus, the augmentations modify the non-causal factors to simulate novel domains. The augmented datasets are realized by pairing the input with the corresponding label and are represented as $(x_{s_i} = \mathcal{T}_i(x_s), y_s) \in \mathcal{D}_{s_i}\ \forall\, i \in [K]$.

Effect of number of source domains K. Having a very high K would lead to significant overlap of the the *leave-one-out* subspaces with the ERM, *i.e.* nullify the advantage of *LO++*. Further, a high K would induce a higher computational complexity. Thus, it becomes essential to filter out augmentations through a principled selection criteria.

Definition 2 (Augmentation selection criteria) *Using Eq. 3, an augmentation $\mathcal{T}_i$ will be selected if $|\gamma_i| < 1$. We give a tractable surrogate for this condition, using a hypothesis $h_s = \arg\min_{h \in \mathcal{A}} \hat{\epsilon}_{(x,y) \in \mathcal{D}_s}(h)$ trained only on $\mathcal{D}_s$,*

$$\hat{\epsilon}_{(\mathcal{T}_i(x),y) \in \mathcal{D}_{s_i}}(h_s) - \hat{\epsilon}_{(x,y) \in \mathcal{D}_s}(h_s) > \tau; \tag{4}$$

i.e. the gap between the empirical risks (i.e. $\hat{\epsilon}$) of h_s on $\mathcal{D}_{s_i}$ and $\mathcal{D}_s$ should be greater than a threshold τ. This ensures that $\mathcal{T}_i$ exerts a substantial alteration in the image statistics equivalent to the style gap between two diverse domains.

Intuitively, an augmentation is selected if it can suppress (*i.e.* $|\gamma_i| < 1$) the original domain factor f_s. In practice, γ_i is intractable in the absence of disentangled f_y and f_s. Thus, we rely on Eq. 4 whose LHS expresses the gener-

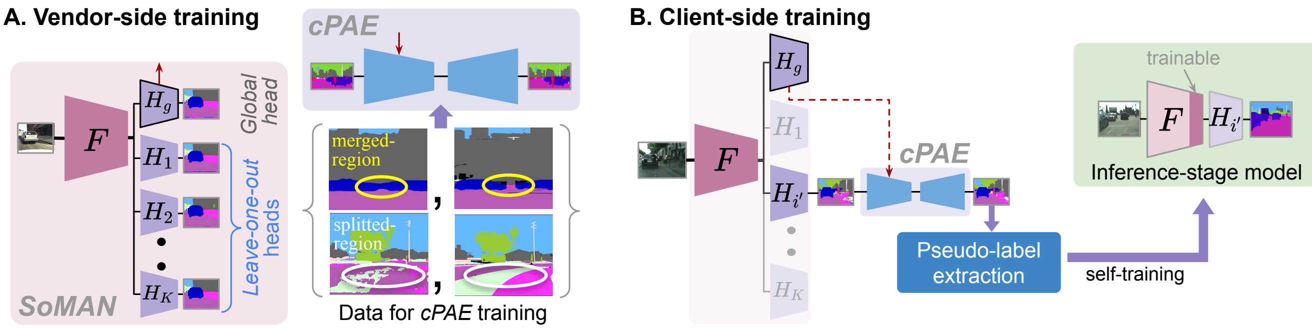

Figure 3. **A.** SoMAN constitutes of a global-head with multiple leave-one-out heads (left). Vendor also trains cPAE to discourage prediction irregularities. **B.** Client receives SoMAN and cPAE from vendor, and extracts robust and meaningful pseudo-labels for target samples via the optimal head $H_{i'}$ to perform *source-free* DA. The inference model uses only the optimal head $H_{i'}$ (no computational overhead).

alization error due to the domain-specific bias (*i.e*. the correlation between f_s and y_s) inculcated in h_s. See Suppl. for an extended explanation of the selection criteria. These diverse domains will help the model generalize to a wider range of targets. Henceforth, we denote each of these as an AG (augmentation-group), each representing a specific type of class-preserving, domain-varying augmentation.

3.1.3 Vendor-side architecture and training

Architecture. Considering the advantages of *LO*++, we propose a Source-only Multi-Augmentation Network, SoMAN, which is essentially a multi-head architecture with a shared CNN backbone F (see Fig. 3A). Along with a global output head H_g which is optimized using ERM, we employ leave-one-out heads $\{H_i\}_{i=1}^K$ trained to be sensitive towards the corresponding AG (*i.e*. $\mathcal{T}_i$) while being invariant to others. Formally, the global head is trained using all the augmented datasets *i.e*. $\mathcal{D}_{s_g} = \cup_{i=1}^K \mathcal{D}_{s_i}$ and each non-global head H_i is trained using a head-specific dataset $\mathcal{D}_{s_i}^- = \mathcal{D}_{s_g} \setminus \mathcal{D}_{s_i}$.

Training procedure. The SoMAN architecture is trained by simultaneously optimizing the spatial segmentation losses computed at the end of each output head. This encourages F to extract a rich multi-source representation which retains domain-sensitive cues (as a result of the leave-one-out setup) alongside the extraction of domain-generic features. We denote the output of global head as $h_g = H_g(F(x))$. Following a similar convention, output of the leave-one-out heads are denoted by $h_i = H_i(F(x))$. Thus, the final objective for end-to-end training of SoMAN is formulated as,

$$\min_{\theta} \sum_{i=1}^{K} \underset{(x,y)\in\mathcal{D}_{s_i}^-}{\mathbb{E}} [-\langle y, \log h_i\rangle] + \underset{(x,y)\in\mathcal{D}_{s_g}}{\mathbb{E}} [-\langle y, \log h_g\rangle] \tag{5}$$

Here, θ denotes a set of parameters from all the heads, *i.e*. $\theta_F, \theta_{H_g}, \{\theta_{H_i}\}_{i=1}^K$ while $\langle .,. \rangle$ represents the dot product of the two inputs. In practice, the expectations are computed by sampling mini-batches from the corresponding datasets.

3.1.4 Conditional prior-enforcing autoencoder (cPAE)

In dense prediction tasks such as semantic segmentation, not all predictions are equally likely. Though the target annotations are not available during the client-side training, we aim to explicitly impart the general knowledge of scene prior to constrain the solution space. The use of scene prior would encourage plausible scene segments while discouraging irregularities (see Fig. 3A) such as “car flying in the sky”, “grass on road”, “split car shape”, “merged pedestrians”, etc. We recognize that the SoMAN may lack the ability to capture the above discussed inductive bias.

How can structural inductive bias be captured? We propose a conditional Prior-enforcing Auto-Encoder (cPAE), denoted by Q, that refines the predicted segmentation maps (seg-maps) conditioned on domain-generic features extracted from SoMAN. Instead of training it as a plain auto-encoder, we plan to train it as a denoising auto-encoder. The question that arises here is: how do we simulate noise for the cPAE inputs? We take advantage of sensitivity of leave-one-out heads to the corresponding AGs to simulate noisy seg-maps. Thus, the cPAE output distribution is $Q(y|F_g(x_{s_i}), \hat{y})$ where $\hat{y} = H_i(F(x_{s_i}))$. F_g consists of the backbone F and the first block of H_g such that $F_g(x_{s_i})$ are domain-generic features since H_g is trained using all AGs. We train the cPAE to align its output distribution with the true source label distribution p_s as follows

$$\min_{\theta_Q} \sum_{i=1}^{K} \mathbb{E}_{(x,y)\in\mathcal{D}_{s_i}} [\text{KL}(p_s(y), Q(y|F_g(x), \hat{y}))] \tag{6}$$

Here, KL indicates the Kullback-Leibler divergence. In practice, cross-entropy loss between the cPAE output and ground truth seg-map is used, derived from the KL term.

3.2. Client-side Strategy

Since the client can access only unlabeled target data $x^t \in \mathcal{D}^t$, we propose the use of self-training for this source-free adaptation step. However, this presents two caveats,

a) Risk of overfitting to wrong overconfident predictions. To counter this, we propose to utilize the multiple heads of

SoMAN and the cPAE to generate reliable pseudo-labels.

b) Loss of task-relevant information. To avoid this, we aim to preserve the task-specific knowledge of the vendor model. While prior arts trained the entire model, we propose to train only a handful of weights belonging to the later layers of F while others are frozen from vendor-side. The frozen output heads hold useful, domain-generic, task-related inductive bias. It also constrains the optimization to operate within the hypothesis subspace of the vendor-side initialization. Thus, the client can leverage the vendor's foresighted preparation to avoid sub-optimal solutions.

3.2.1 Pseudo-label extraction via cPAE

Since pseudo-labels are the only supervision signal in the proposed *source-free* self-training, it is crucial to ensure that they are highly informative and reliable. To this end, we propose to utilize the optimal head of the vendor provided SoMAN and the cPAE to obtain improved pseudo-labels. We consider the optimal head as the one that produces the lowest average self-entropy for the target training dataset. Formally, $H_{i'}$ is the optimal head where $i' = \arg\min_{i \in \{g,[K]\}} \sum_{x \in \mathcal{D}_t} \{-\langle h_i, \log h_i \rangle\}$ where $h_i = H_i(F(x))$. The optimal prediction can be represented as $Q(h_{i'})$. Note that we denote the cPAE output as $Q(h_{i'})$ omitting the conditional feature input for simplicity.

Using the optimal prediction, we follow [41, 95] for the confidence thresholding method. Particularly, we choose the top 33% of the most confident pixel-level predictions per class over the entire target training set. This gives a target pseudo-labeled subset $(x_t, \hat{y}_t) \in \hat{\mathcal{D}}_t$ for self-training. Note that, the unselected pixels are assigned a separate, 'unknown' class which is not considered in training.

3.2.2 Source-free adaptation via self-training

We perform three rounds of self-training, following [83], where each round consists of pseudo-label extraction in an offline manner followed by supervised training on the extracted pseudo-labels. Entropy minimization is used as a regularizer during self-training. Further, we use the shared backbone F along with the optimal head, $H_{i'}$, for both self-training and test-time inference. Formally,

$$\min_{\theta_F} \mathop{\mathbb{E}}_{(x_t, \hat{y}_t) \in \hat{\mathcal{D}}_t} [-\langle \hat{y}_t, \log H_{i'}(F(x_t)) \rangle] \tag{7}$$

3.2.3 Test-time inference

As we propose only optimal head (*i.e.* $H_{i'}$) self-training, our inference-stage model is $H_{i'}(F(x_t))$ as shown in Fig. 3B. However, cPAE provides a further improvement in performance if used during inference. But, unless otherwise specified, the experiments use only $H_{i'}(F(x_t))$ for self-training and evaluation, for a fair comparison. Note that, 'w/ cPAE' means that cPAE was used only for pseudo-label extraction.

4. Experiments

We perform a thorough evaluation of our approach against *state-of-the-art* prior works across multiple settings.

4.1. Experimental Settings

a) Network architectures. Following [41, 83], we employ 2 widely-used network architectures for the DA setting on semantic segmentation, DeepLabv2 [4] with ResNet101 [18] backbone and FCN8s [45] with VGG16 [62] backbone. See Suppl. for the complete details.

b) Datasets. We extensively evaluate the proposed approach on two popular synthetic-to-real benchmarks *i.e.* GTA5→Cityscapes and SYNTHIA→Cityscapes. We provide the complete implementation details in the Suppl.

c) Evaluation metric. Following [41, 83], we compute per-class IoUs as well as mean IoU (mIoU) over all 19 classes for the GTA5→Cityscapes task. For SYNTHIA→Cityscapes, we report the same for 13 and 16 classes because of the lower number of overlapping classes. Following [50, 74, 88], we use multi-scale testing. Due to space limitations, we report mean IoUs for *class-groups*[2] instead of reporting IoUs for each individual class.

d) Augmentations. We select the following $K = 5$ AGs (see Fig. 2D) using Definition 2 with the mIoU metric.

Aug-A (FDA [83]): This uses Fourier transform to transfer style from a reference image while retaining the semantic features [82] of the input. While FDA [83] transfers the style from target images, we do not access target data for vendor-side training. We use a small subset from style transfer dataset [23] and random noise as reference images.

Aug-B (Style augmentation [25]): This technique uses a deep style transfer network for style randomization by randomly sampling a style embedding from a multivariate normal distribution instead of using reference style image. This provides practically infinite number of stylization options.

Aug-C (AdaIN [23]): This uses Adaptive Instance Normalization (AdaIN) layers to inject style from a given reference image. In contrast to Aug-B, this provides a way to stylize images using a desired style image. We use a small subset from style transfer dataset [23] as reference images.

Aug-D (Weather augmentation) [27, 51]: We use realistic weather augmentations to generate varying levels of snow and frost in the images. Compared to other AGs, this simulates realistic variations in the road scene images.

Aug-E (Cartoon augmentation) [27]: This technique generates cartoonized versions of input images. This augmentation is diverse and useful as it produces almost texture-less images as in cartoons or comic books.

[2]Background (**BG**) - building, wall, fence, vegetation, terrain, sky; Minority Class (**MC**) - rider, train, motorcycle, bicycle; Road Infrastructure Vertical (**RIV**) - pole, traffic light, traffic sign; Road Infrastructure Ground (**RIG**) - road, sidewalk; and Dynamic Stuff (**DS**) - person, car, truck, bus.

Table 1. Quantitative evaluation on GTA5→Cityscapes. Performance on different segmentation architectures: A (DeepLabv2 ResNet-101), B (FCN8s VGG-16). SF indicates *source-free* adaptation. See Suppl. for the extended table with per-class IoUs. *Ours (V)* indicates use of our vendor-side `AG`s with prior art, * indicates results produced using the released code of prior arts.

#	Method	Arch.	SF	BG	MC	RIV	RIG	DS	mIoU
1.	PLCA [28]	A	×	57.3	28.3	31.1	57.2	60.2	47.7
2.	CrCDA [22]	A	×	57.5	24.5	33.8	73.9	57.6	48.6
3.	RPT [88]	A	×	62.5	34.9	42.0	67.3	59.4	53.2
4.	DACS [67]	A	×	63.1	24.2	45.9	64.7	61.8	52.1
5.	FADA [74]	A	×	61.9	26.7	35.0	70.8	56.7	50.1
6.	IAST [50]	A	×	60.4	32.6	34.1	76.5	60.7	52.2
7.	*Ours (V)* + FADA*	A	×	62.8	27.1	35.3	71.1	57.2	50.6
8.	*Ours (V)* + IAST*	A	×	61.0	33.1	34.6	77.1	61.2	52.8
9.	URMA [63]	A	✓	55.8	23.8	22.3	73.7	52.8	45.1
10.	SRDA* [2]	A	✓	57.1	20.2	33.5	68.8	51.9	45.8
11.	*Ours (w/o `cPAE`)*	A	✓	61.8	30.3	35.1	69.2	60.8	51.6
12.	*Ours (w/ `cPAE`)*	A	✓	62.8	33.4	36.2	72.0	66.4	**53.4**
13.	LTIR [29]	B	×	58.6	14.0	26.5	73.5	42.5	42.3
14.	FADA [74]	B	×	57.7	16.3	25.8	71.7	50.1	43.8
15.	PCEDA [82]	B	×	56.4	20.5	31.2	67.5	49.5	44.6
16.	SFDA [44]	B	✓	51.6	7.8	15.9	58.6	43.7	35.8
17.	*Ours (w/o `cPAE`)*	B	✓	54.7	19.9	27.3	66.2	50.3	43.4
18.	*Ours (w/ `cPAE`)*	B	✓	49.9	30.3	32.9	74.9	50.8	**45.9**

Table 2. Ablation study for GTA5→Cityscapes. * indicates 3 rounds of self-training after the mentioned method. The client-side ablations begin from the best vendor-side model.

	Method	mIoU
Vendor-side	Standard single-source*	44.4
	Multi-source ERM*	47.6
	Domain-experts++ (DE++)*	48.0
	Leave-one-out++ (LO++)*	**51.6**
Client-side	w/o `cPAE`	51.6
	+ Inference via `cPAE`	52.5
	w/ `cPAE`	53.4
	+ Inference via `cPAE`	**54.2**

4.2. Discussion

We provide an extensive ablation study of both the *vendor-side* and the *client-side* preparation. Further, we show that our approach generalizes across novel target scenarios and is compatible to online domain adaptation.

4.2.1 Comparison with prior arts.

We compare our proposed approach with prior arts in Table 1 and 3. We also compare our vendor-side approach with prior DG works in Table 4. Our method achieves *state-of-the-art* performance across all benchmarks. We also present the qualitative evaluation of our approach in Fig. 4.

Our proposed client-side adaptation is more scalable compared to prior works like PCEDA [82], RPT [88], IAST [50] in two ways. First, our method does not require image-to-image translation networks (PCEDA) or adversarial training (RPT, IAST) thereby reducing the adaptation complexity. Also note that the frozen `cPAE` is used only to obtain better pseudo-labels and is not involved in backpropagation for adaptation training. Second, the client can perform adaptation to multiple different target domains without the complex vendor-side training and without access to the source data. We study the second aspect further in the paper.

Table 3. Quantitative evaluation on SYNTHIA→Cityscapes. Performance on different segmentation architectures: A (DeepLabv2 ResNet-101), B (FCN8s VGG-16). mIoU and mIoU* are averaged over 16 and 13 categories respectively. SF indicates whether the method supports *source-free* adaptation. See Suppl. for the extended table with per-class IoUs.

#	Method	Arch.	SF	BG	MC	RIV	RIG	DS	mIoU	mIoU*
1.	CAG [86]	A	×	81.3	32.9	18.0	62.6	54.9	44.5	52.6
2.	USAMR [91]	A	×	81.3	33.0	25.1	60.7	61.7	46.5	53.8
3.	DACS [67]	A	×	85.4	38.1	23.3	52.8	63.1	48.7	54.8
4.	RPL [90]	A	×	81.8	32.8	25.6	64.8	63.3	-	54.9
5.	IAST [50]	A	×	83.9	38.9	29.9	61.7	63.4	49.8	57.0
6.	RPT [88]	A	×	85.7	37.2	35.1	68.2	66.2	51.7	59.5
7.	URMA [63]	A	✓	80.1	23.6	25.1	41.9	46.6	39.6	45.0
8.	*Ours (w/o `cPAE`)*	A	✓	82.9	34.4	22.5	66.8	65.3	48.1	55.5
9.	*Ours (w/ `cPAE`)*	A	✓	84.3	42.2	29.3	69.8	67.8	**52.0**	**60.1**
10.	PyCDA [42]	B	×	75.4	16.4	24.0	53.6	47.6	35.9	42.6
11.	SD [14]	B	×	79.2	6.3	10.7	64.4	54.4	-	43.4
12.	FADA [74]	B	×	82.1	16.1	15.1	58.2	52.6	39.5	46.0
13.	BDL [41]	B	×	78.3	25.2	17.7	51.2	50.5	39.0	46.1
14.	PCEDA [82]	B	×	79.8	30.7	19.5	57.5	49.2	41.1	48.7
15.	*Ours (w/o `cPAE`)*	B	✓	82.0	9.5	21.9	67.0	51.4	40.0	46.7
16.	*Ours (w/ `cPAE`)*	B	✓	83.1	17.7	24.5	69.4	51.8	**41.3**	**48.9**

Table 4. Domain generalization evaluation. For SYNTHIA, mIoU computed over 16 categories. SO, ERM and LO indicate *source-only*, empirical risk minimization and *leave-one-out* respectively.

Method	GTA5→Cityscapes		SYNTHIA→Cityscapes	
	ResNet-101	VGG16	ResNet-101	VGG16
IBN-Net [54]	37.1	34.7	35.6	33.0
ASG [6]	38.8	35.4	36.9	34.2
DRPC [85]	42.5	-	37.6	35.5
Ours (ERM) (SO)	43.1	38.9	40.1	36.9
Ours (LO++) (SO)	**43.5**	**39.2**	**40.6**	**37.4**

a) Comparison with source-free prior arts. We implemented [2] for GTA5→Cityscapes (see #10-12 in Table 1) since they only report results for single object segmentation. We outperform their approach by a significant margin (8.1%). We also compare with concurrent source-free works [44, 63] (see #9 vs. #12, #16 vs. #18 in Table 1 and #7 vs. #9 in Table 3) and outperform them by $\sim$12%.

b) Disentangling the gains from use of augmented data. We show the results for 2 prior arts [50, 74] using our vendor-side `AG`s during training (#5-8 in Table 1). While the performance improves compared to that originally reported, our proposed method (#12 in Table 1) still outperforms them. Thus, the improvement of our proposed method depends not only on the use of `AG`s but also on the multi-head, leave-one-out `SoMAN` framework and the `cPAE`.

4.2.2 Ablation study

Table 2 reports a detailed ablation to independently analyse the components of the vendor and client side strategies.

First, we evaluate the effectiveness of the proposed vendor-side strategies. For a fair comparison, we use a consistent client-side training for all the vendor-side ablations. As a baseline, we employ a standard (unaugmented) single-

Table 5. Evaluating generalization and compatibility to online adaptation for GTA5→Cityscapes models on Foggy-Cityscapes and NTHU-Cross-City datasets. 0.005, 0.01, and 0.02 indicate the levels of fog in the dataset and GT indicates ground truth segmentation maps. * indicates experiment reproduced by us using the released code of prior arts. We also show standard Cityscapes results for reference.

	#	Method	Access to GTA5 | Citysc.	Cityscapes		Foggy-Cityscapes (19-class)				NTHU-Cross-City (13-class)				
				19-class	13-class	0.005	0.01	0.02	Avg.	Rio	Rome	Taipei	Tokyo	Avg.
Vendor-side (GTA5)	1.	BDL (w/o ST) [41]	✓ | ✓(no GT)	43.3	53.2	40.4	36.8	30.3	35.8	38.9	42.2	42.2	41.2	41.1
	2.	FDA* (w/o ST) [83]	✓ | ✓(no GT)	42.7	51.9	42.1	40.3	35.3	39.2	42.2	42.3	37.5	42.3	41.0
	3.	*Ours (vendor-side)*	✓ | ×	43.1	51.5	43.6	42.4	38.3	**41.4**	47.0	48.7	43.4	44.5	**45.9**
Client-side (→Citysc.)	4.	ASN [68]	✓ | ✓(no GT)	42.4	51.1	41.0	38.0	31.7	36.9	41.8	44.5	37.5	41.9	41.4
	5.	MSL [5]	✓ | ✓(no GT)	46.4	54.5	44.3	40.9	34.2	39.8	44.4	47.0	45.6	44.7	45.4
	6.	BDL [41]	✓ | ✓(no GT)	48.5	57.7	46.0	42.6	36.3	41.6	44.1	47.1	47.5	44.3	45.7
	7.	FDA [83]	✓ | ✓(no GT)	48.8	57.8	47.6	45.2	39.1	44.0	47.8	46.6	42.7	48.1	46.3
	8.	*Ours (client-side)*	× | ✓(no GT)	53.4	61.4	51.7	48.9	42.3	**47.6**	47.1	47.7	45.7	46.5	**46.7**
Online Adapt. (→FoggyC / →NTHU)	9.	CBST [95]	× | ✓(w/ GT)	-	-	-	-	-	-	52.2	53.6	50.3	48.8	51.2
	10.	MSL [5]	× | ✓(w/ GT)	-	-	-	-	-	-	53.3	54.5	50.6	50.5	52.2
	11.	CSCL [13]	× | ✓(w/ GT)	-	-	-	-	-	-	53.8	54.8	51.4	51.0	52.7
	12.	*Ours (client-side)*	× | ×	-	-	53.6	51.1	45.9	50.2	54.3	55.0	51.6	51.3	**53.0**

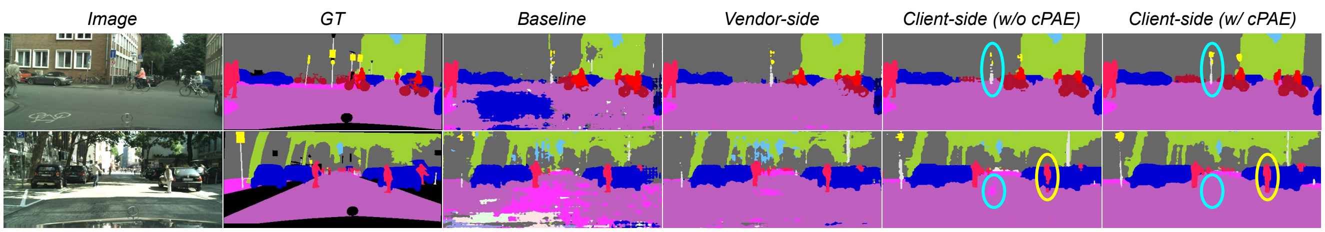

Figure 4. Qualitative evaluation of the proposed approach. Vendor-side model generalizes better than the baseline but performs worse than client-side due to the domain gap. Inculcating the prior knowledge from cPAE structurally regularizes the predictions and overcomes the merged-region (yellow circle) and split-region (blue circle) problems. See Suppl. for extended evaluation. *Best viewed in color.*

source-trained model. The ERM model gives an improvement of 3.2% over the baseline. Next, we evaluate *DE++* and observe an improvement of 0.4%. *LO++* gives a further improvement of 3.6% over *DE++*. This shows the clear superiority of *LO++* over both ERM and *DE++*.

Second, under client-side ablation, cPAE for pseudo-label extraction gives a boost of 1.8%. Further, using cPAE for inference gives an additional 0.8-0.9% improvement.

4.2.3 Analyzing cross-dataset generalization

Unlike prior arts which assume concurrent access to source and target (inculcates target-bias), our target-free vendor-side model is expected to generalize well to unseen targets. To this end, Table 5 shows our generalizability to other road-scene datasets, such as Foggy-Cityscapes [59] and NTHU-Cross-City [10]; before (#1-3) and after (#4-8) self-training on the related real domain, *i.e*. Cityscapes. Among different variants, we achieve a superior average generalization even without concurrent access to samples from the related domain, Cityscapes. Note that concurrent access is beneficial to better characterize the domain gap.

4.2.4 Compatibility to online domain adaptation

Online adaptation [26, 49] refers to a deployment setting where the model is required to continuously adapt to the current working conditions. The current state of the model may overcome its past domain-biases to perform the best at a given scenario. The proposed *client-side* training can be seen as an online adaptation algorithm. Here, the frozen parameters of the multi-head SoMAN helps to retain task-specific knowledge while allowing adaptation to unlabeled samples from the new environment. In the last section of Table 5, the initial Cityscapes adapted SoMAN is independently adapted to different secondary domains under Foggy-Cityscapes and NTHU-Cross-City. We also compare our results with recent Cityscapes→NTHU-Cross-City works (#9-12) that concurrently access labeled Cityscapes and unlabeled NTHU-Cross-City datasets. The improved performance shows our compatibility to online adaptation.

5. Conclusion

We introduced a *source-free* DA framework for semantic segmentation, recognizing practical scenarios where source and target data are not concurrently accessible. We cast the *vendor-side* training as multi-source learning. Based on theoretical insights, we proposed SoMAN that balances generalization and specificity using the systematically selected AGs without access to the target. To provide a strong support for the dense prediction task, cPAE is trained to denoise segmentation predictions and improve pseudo-label quality for *client-side* source-free self-training. Extending this framework to more DA scenarios involving category-shift can be a direction for future research.

Acknowledgements. This work was supported by MeitY (Ministry of Electronics and Information Technology) project (No. 4(16)2019-ITEA), Govt. of India.

References

[1] Surbhi Aggarwal, Jogendra Nath Kundu, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Wamda: Weighted alignment of sources for multi-source domain adaptation. In *BMVC*, 2020. 1

[2] Mathilde Bateson, Hoel Kervadec, Jose Dolz, Herve Lombaert, and Ismail Ben Ayed. Source-relaxed domain adaptation for image segmentation. In *MICCAI*, 2020. 2, 7

[3] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *CVPR*, 2019. 2

[4] 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. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 6

[5] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *ICCV*, 2019. 3, 8

[6] Wuyang Chen, Zhiding Yu, Zhangyang Wang, and Anima Anandkumar. Automated synthetic-to-real generalization. In *ICML*, 2020. 3, 7

[7] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *CVPR*, 2019. 2

[8] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *CVPR*, 2018. 2

[9] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *CVPR*, 2019. 2

[10] 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. 1, 8

[11] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *ICCV*, 2019. 2, 3

[12] 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. 2

[13] Jiahua Dong, Yang Cong, Gan Sun, Yuyang Liu, and Xiaowei Xu. Cscl: Critical semantic-consistent learning for unsupervised domain adaptation. In *ECCV*, 2020. 2, 8

[14] Liang Du, Jingang Tan, Hongye Yang, Jianfeng Feng, Xiangyang Xue, Qibao Zheng, Xiaoqing Ye, and Xiaolin Zhang. Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation. In *ICCV*, 2019. 2, 7

[15] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *CVPR*, 2019. 2

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[17] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *NeurIPS*, 2005. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[19] 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. 2

[20] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. FCNs in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016. 2

[21] Haoshuo Huang, Qixing Huang, and Philipp Krahenbuhl. Domain transfer through deep activation matching. In *ECCV*, 2018. 2

[22] Jiaxing Huang, Shijian Lu, Dayan Guan, and Xiaobing Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. In *ECCV*, 2020. 7

[23] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 6

[24] Maximilian Ilse, Jakub M Tomczak, and Patrick Forré. Selecting data augmentation for simulating interventions. In *ICML*, 2021. 1, 2

[25] Philip T Jackson, Amir Atapour-Abarghouei, Stephen Bonner, Toby P Breckon, and Boguslaw Obara. Style augmentation: Data augmentation via style randomization. In *CVPR Workshops*, 2019. 6

[26] Vidit Jain and Erik Learned-Miller. Online domain adaptation of a pre-trained cascade of classifiers. In *CVPR*, 2011. 8

[27] Alexander B. Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kilian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al. imgaug. https://github.com/aleju/imgaug, 2020. Online; accessed 01-Feb-2020. 6

[28] Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, and Alexander Hauptmann. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. In *NeurIPS*, 2020. 7

[29] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *CVPR*, 2020. 7

[30] Minsu Kim, Sunghun Joung, Seungryong Kim, JungIn Park, Ig-Jae Kim, and Kwanghoon Sohn. Cross-domain grouping and alignment for domain adaptive semantic segmentation. In *AAAI*, 2021. 2

[31] Jogendra Nath Kundu, Phani Krishna Uppala, Anuj Pahuja, and R. Venkatesh Babu. Adadepth: Unsupervised content

congruent adaptation for depth estimation. In *CVPR*, 2018. 2

[32] Jogendra Nath Kundu, Nishank Lakkakula, and R. Venkatesh Babu. Um-adapt: Unsupervised multi-task adaptation using adversarial cross-task distillation. In *ICCV*, 2019. 2

[33] Jogendra Nath Kundu, Gaurav Singh Rajput, and Venkatesh Babu RADHAKRISHNAN. Vrt-net: Real-time scene parsing via variable resolution transform. In *WACV*, 2020. 2

[34] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *CVPR*, 2020. 1

[35] Jogendra Nath Kundu, Naveen Venkat, Ambareesh Revanur, Rahul M V, and R. Venkatesh Babu. Towards inheritable models for open-set domain adaptation. In *CVPR*, 2020. 3

[36] Jogendra Nath Kundu, Rahul Mysore Venkatesh, Naveen Venkat, Ambareesh Revanur, and R. Venkatesh Babu. Class-incremental domain adaptation. In *ECCV*, 2020. 3

[37] Suhyeon Lee, Junhyuk Hyun, Hongje Seong, and Euntai Kim. Unsupervised domain adaptation for semantic segmentation by content transfer. In *AAAI*, 2021. 2

[38] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017. 1, 3

[39] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. Content-consistent matching for domain adaptive semantic segmentation. In *ECCV*, 2020. 3

[40] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *CVPR*, 2020. 1

[41] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*, 2019. 2, 6, 7, 8

[42] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *ICCV*, 2019. 7

[43] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 2020. 1

[44] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *CVPR*, 2021. 2, 7

[45] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 6

[46] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *ICCV*, 2019. 2

[47] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Adversarial style mining for one-shot unsupervised domain adaptation. In *NeurIPS*, 2020. 2

[48] 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. In *CVPR*, 2019. 2

[49] Massimiliano Mancini, Hakan Karaoguz, Elisa Ricci, Patric Jensfelt, and Barbara Caputo. Kitting in the wild through online domain adaptation. In *IROS*, 2018. 8

[50] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *ECCV*, 2020. 3, 6, 7

[51] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 6

[52] Luigi Musto and Andrea Zinelli. Semantically adaptive image-to-image translation for domain adaptation of semantic segmentation. In *BMVC*, 2020. 2

[53] 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. 3

[54] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*, 2018. 3, 7

[55] Sujoy Paul, Yi-Hsuan Tsai, Samuel Schulter, Amit K. Roy-Chowdhury, and Manmohan Chandraker. Domain adaptive semantic segmentation using weak labels. In *ECCV*, 2020. 3

[56] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *ICML*, 2020. 4

[57] Fabio Pizzati, Raoul de Charette, Michela Zaccaria, and Pietro Cerri. Domain bridge for unpaired image-to-image translation and unsupervised domain adaptation. In *WACV*, 2020. 2

[58] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018. 1

[59] 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. 1, 8

[60] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *CVPR*, 2018. 2

[61] Inkyu Shin, Sanghyun Woo, Fei Pan, and In So Kweon. Two-phase pseudo label densification for self-training based domain adaptation. In *ECCV*, 2020. 3

[62] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 6

[63] Prabhu Teja Sivaprasad and Francois Fleuret. Uncertainty reduction for model adaptation in semantic segmentation. In *CVPR*, 2021. 2, 7

[64] M Naseer Subhani and Mohsen Ali. Learning from scale-invariant examples for domain adaptation in semantic segmentation. In *ECCV*, 2020. 3

[65] Marco Toldo, Andrea Maracani, Umberto Michieli, and Pietro Zanuttigh. Unsupervised domain adaptation in semantic segmentation: A review. *Technologies*, 8(2), 2020. 2

[66] Marco Toldo, Umberto Michieli, and Pietro Zanuttigh. Unsupervised domain adaptation in semantic segmentation via orthogonal and clustered embeddings. In *WACV*, 2021. 2

[67] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. DACS: Domain adaptation via cross-domain mixed sampling. In *WACV*, 2021. 7

[68] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018. 8

[69] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *ICCV*, 2019. 2

[70] Vladimir Vapnik. Principles of risk minimization for learning theory. In *NeurIPS*, 1992. 3

[71] Naveen Venkat, Jogendra Nath Kundu, Durgesh Kumar Singh, Ambareesh Revanur, and R. Venkatesh Babu. Your classifier can secretly suffice multi-source domain adaptation. In *NeurIPS*, 2020. 1

[72] 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. 2

[73] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. DADA: Depth-aware domain adaptation in semantic segmentation. In *ICCV*, 2019. 2

[74] Haoran Wang, Tong Shen, Wei Zhang, Lingyu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *ECCV*, 2020. 2, 6, 7

[75] Kaihong Wang, Chenhongyi Yang, and Margrit Betke. Consistency regularization with high-dimensional non-adversarial source-guided perturbation for unsupervised domain adaptation in segmentation. In *AAAI*, 2021. 2

[76] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *CVPR*, 2020. 2

[77] 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. 2

[78] Zuxuan Wu, Xin Wang, Joseph E Gonzalez, Tom Goldstein, and Larry S Davis. Ace: adapting to changing environments for semantic segmentation. In *ICCV*, 2019. 2

[79] Jinyu Yang, Weizhi An, Sheng Wang, Xinliang Zhu, Chaochao Yan, and Junzhou Huang. Label-driven reconstruction for domain adaptation in semantic segmentation. In *ECCV*, 2020. 2

[80] Jinyu Yang, Weizhi An, Chaochao Yan, Peilin Zhao, and Junzhou Huang. Context-aware domain adaptation in semantic segmentation. In *WACV*, 2021. 2

[81] Jihan Yang, Ruijia Xu, Ruiyu Li, Xiaojuan Qi, Xiaoyong Shen, Guanbin Li, and Liang Lin. An adversarial perturbation oriented domain adaptation approach for semantic segmentation. In *AAAI*, 2020. 2

[82] Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. Phase consistent ecological domain adaptation. In *CVPR*, 2020. 6, 7

[83] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *CVPR*, 2020. 2, 3, 6, 8

[84] Fei Yu, Mo Zhang, Hexin Dong, Sheng Hu, Bin Dong, and Li Zhang. DAST: Unsupervised domain adaptation in semantic segmentation based on discriminator attention and self-training. In *AAAI*, 2021. 2

[85] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *ICCV*, 2019. 3, 7

[86] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. In *NeurIPS*, 2019. 7

[87] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *CVPR*, 2018. 2

[88] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Dong Liu, and Tao Mei. Transferring and regularizing prediction for semantic segmentation. In *CVPR*, 2020. 6, 7

[89] Sicheng Zhao, Bo Li, Xiangyu Yue, Yang Gu, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source domain adaptation for semantic segmentation. In *NeurIPS*, 2019. 1, 3

[90] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision (IJCV)*, 2020. 3, 7

[91] Zhedong Zheng and Yi Yang. Unsupervised scene adaptation with memory regularization in vivo. In *IJCAI*, 2020. 3, 7

[92] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2

[93] Xinge Zhu, Hui Zhou, Ceyuan Yang, Jianping Shi, and Dahua Lin. Penalizing top performers: Conservative loss for semantic segmentation adaptation. In *ECCV*, 2018. 2

[94] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *ICCV*, 2019. 2

[95] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018. 2, 6, 8