SOURCE-TARGET COORDINATED TRAINING WITH MULTI-HEAD HYBRID-ATTENTION FOR DOMAIN ADAPTIVE SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Domain adaptive semantic segmentation aims to densely assign semantic labels for each pixel on the unlabeled target domain by transferring knowledge from the labeled source domain. Due to the domain shift problem, the success of adaptation on the unseen domain depends on the feature alignment between different domains. Hence, this paper focuses on feature alignment for domain adaptive semantic segmentation, *i.e.*, when to align and how to align. Since no label is available in the target domain, aligning the target distribution too early would lead to poor performance due to pseudo-label noise, while too late may cause the model to underfit the target domain. In this paper, we propose a Source-Target Coordinated Training (STCT) framework, where a coordination weight is designed to control the time to align. For the problem of how to align, we design a Multi-head Hybrid-Attention (MHA) module to replace the multi-head self-attention (MSA) module in the transformer. The proposed MHA module consists of intra-domain self-attention and inter-domain cross-attention mechanisms. Compared with the MSA module, the MHA module achieves feature alignment by explicitly constructing interaction between different domains without additional computations and parameters. Moreover, to fully explore the potential of the proposed MHA module, we comprehensively investigate different designs for the MHA module and find some important strategies for effective feature alignment. Our proposed method achieves competitive performance on two challenging synthetic-to-real benchmarks, GTA5-to-CityScapes and SYNTHIA-to-Cityscapes.

1 INTRODUCTION

Deep neural networks have achieved remarkable success in various application scenarios, but it still suffers from expensive human-labour annotation and poor adaptation performance. Thus, as a promising technique, unsupervised domain adaptation attracts much attention from academia and industry, especially for dense prediction tasks. Unsupervised domain adaptation for semantic segmentation is proposed to make semantic predictions for each pixel on the unlabeled target domain by learning a model with labeled source domain images. However, due to the significant distribution discrepancy between different domains, *i.e.*, the domain shift problem, the model trained on the source domain shows a remarkable performance drop on the target domain.

Numerous methods are proposed to achieve feature alignment by learning domain-invariant features to address the domain shift problem. Pixel-level alignment methods Li et al. (2019); Yang et al. (2020); Kim & Byun (2020); Cheng et al. (2021) utilize an image translation model, such as GAN Zhu et al. (2017), and a segmentation method iteratively to project the image styles of different domains into the same domain. Prototype-level alignment methods Zhang et al. (2021); Liu et al. (2021a) minimize distances between the class prototypes of the source and target domains. Label-level alignment methods Tsai et al. (2018); Vu et al. (2019) exploit the similarity of probability and entropy to produce similar predictive distributions in the output space of the source and target domains. However, there are two issues rarely mentioned by the previous methods. The first problem, is when it is most appropriate to involve target domain data for model training (*i.e.*, "when to align"). Second, most previous approaches learn domain-invariant features implicitly by sharing network pa-

rameters for source and target domains, without explicitly modeling the relationship between source and target domain features (*i.e.*, "how to align"). We aim to solve the two issues step by step.

In domain adaptation, training source and target data together from the beginning does not bring satisfactory results. We argue that how to coordinate the supervised learning process of the source domain with the unsupervised learning of the target domain is crucial for domain adaptive semantic segmentation. Over-training on the source domain prevents the model from learning the domain adaptive features on the target domain. On the contrary, over-training on the target domain results in the model not learning the discriminative category features due to the absence of labels. In addition, training too early on the target domain introduces noisy labels, while training too late traps the model into a local optimum, biased to the source domain feature distribution. Therefore, we take the pseudo-accuracy on target domain as the metric and propose a coordination weight to control the involvement of the target domain in the training process. Based on the coordination weight, we propose a Source-Target Coordinated Training (STCT) framework, an end-to-end self-training framework to coordinate the target domain with the source domain during the training process.

To explicitly exploit the feature correlation between different domains to achieve feature alignment, we propose a Multi-head Hybrid-Attention (MHA) module by incorporating intra-domain self-attention and inter-domain cross-attention mechanisms. Intra-domain self-attention mechanism fully utilizes the label information from the source domain to learn discriminative representations. Inter-domain cross-attention mechanism constructs feature interaction between different domains and fuses features to facilitate feature alignment. Unlike previous domain-separated training methods Zhang et al. (2021); Tsai et al. (2018); Vu et al. (2019); Zou et al. (2018; 2019); Mei et al. (2020); Wang et al. (2021b); Zheng & Yang (2021), our method can be trained on a mixture of source and target features, which facilitates learning domain-invariant features for alignment.

In addition, we design a series of experiments to further explore the potential of the proposed MHA module from two aspects: 1) how to select suitable tokens for alignment in our MHA module; 2) the importance of source and target features in the alignment. From the two aspects, some indispensable component is discovered for effective feature alignment in the proposed MHA module. As a result, we propose a bidirectional semantic-grouping MHA module to fully utilize the capabilities of the inter-domain cross-attention mechanism in feature alignment.

We summarize our contributions as follows. 1) For "when to align" issue, we propose a Source-Target Coordinated Training (STCT) framework based on a coordination weight to achieve balance training between source and target domains. 2) For "how to align" issue, we propose a Multi-head Hybrid-Attention (MHA) module to explicitly construct the feature interaction between different domains and achieve feature alignment. To further explore the potential of MHA, we conduct comprehensive experiments and find some useful strategies to build MHA with outstanding performance. 3) We achieve a comparable performance of 68.05 on GTAV Richter et al. (2016) to Cityscapes Cordts et al. (2016) task and 59.8 on SYNTHIA Ros et al. (2016) to Cityscapes Cordts et al. (2016) task.

2 Related Work

2.1 Domain adaptive semantic segmentation

The main challenge of unsupervised domain adaptive semantic segmentation is the domain shift problem, due to the distribution discrepancy between the source and target domains. Thus, previous works have shown remarkable progress by achieving feature alignment, and can be summarized into the following categories. Pixel-level alignment methods Li et al. (2019); Yang et al. (2020); Kim & Byun (2020); Cheng et al. (2021) first transferred the image style of different domains into the same domain by a style translation model CycleGAN Zhu et al. (2017). Then, a segmentation model is trained on domains with the translated style. Prototype-level alignment methods Zhang et al. (2021); Liu et al. (2021a) utilize the class prototype from the source and target domains to achieve feature alignment. Label-level alignment methods exploited the probability similarity Tsai et al. (2018) and entropy similarity Vu et al. (2019) to generate similar prediction distributions in the output space for either source or target domains. Self-training methods Zou et al. (2018; 2019); Mei et al. (2020); Wang et al. (2021b); Zheng & Yang (2021); Araslanov & Roth (2021) first generate pseudo-labels based on a pre-trained model from the source domain. Then, the model is trained on the target domain with the supervision of pseudo-labels.

2.2 SELF-ATTENTION AND CROSS-ATTENTION MECHANISMS

The self-attention mechanism is the core component of Transformer Vaswani et al. (2017). Many works Han et al. (2020); Dosovitskiy et al. (2020); Liu et al. (2021b) have shown its effectiveness for computer-vision tasks. ViT Dosovitskiy et al. (2020) split an image into feature tokens and took self-attention mechanism to construct relation between feature tokens. Swin Transformer Liu et al. (2021b) introduced the hierarchical structure into ViT Dosovitskiy et al. (2020) and proposed shifted windowing scheme, where self-attention is adopted within local windows for efficient computation. The cross-attention mechanism has shown great potential in feature fusion and feature alignment. Gao Gao et al. (2019) proposed a dynamic fusion with intra-modality and inter-modality attention flow, which exploited the association weights between visual modal and text modal on the visual question answer task. Chen Chen et al. (2021) designed a dual transformer architecture, where the cross-attention mechanism is adopted to exchange information between small-patch and large-patch tokens. Xu Xu et al. (2021) introduced the cross-attention into domain adaptive classification to achieve label denoising. In this paper, we take advantages of cross-attention on feature alignment to build our MHA module for better performance on domain adaptive semantic segmentation.

3 THE PROPOSED METHOD

In this section, we first propose a Source-Target Coordinated Training (STCT) framework, in which a coordination weight is designed to determine the involvement of the target domain in the training progress, as shown in Fig. 1. Then, we propose a Multi-head Hybrid-Attention (MHA) module, integrating intra-domain self-attentive and inter-domain cross-attentive mechanisms to achieve feature alignment, as illustrated in Fig. 2. Finally, we investigate the effects of different designs used for the MHA module. Some strategies are created during the investigation and adopted for our best MHA module, *i.e.*, a bidirectional semantic-grouping MHA.

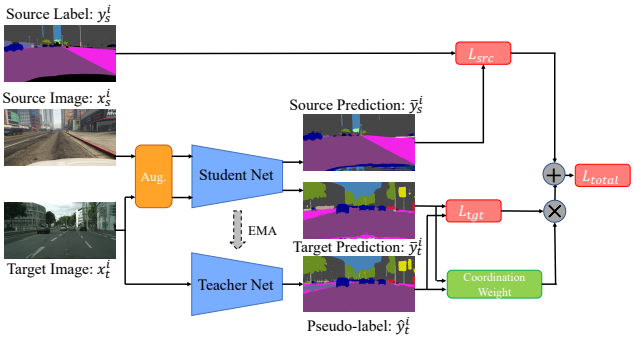


Figure 1: **Illustration of the STCT framework.**

3.1 SOURCE-TARGET COORDINATED TRAINING FRAMEWORK

Given a labeled source domain dataset $\mathbb{D}_s = \{(\boldsymbol{x}_s^i, \boldsymbol{y}_s^i) | \boldsymbol{y}_s^i \in \mathbb{R}^{H \times W}\}_{i=1}^{N_s}$ and an unlabeled target domain dataset $\mathbb{D}_t = \{\boldsymbol{x}_t^i\}_{i=1}^{N_t}$, unsupervised domain adaptive semantic segmentation predicts pixel-level semantic masks for target domain images, where N_s and N_t are the numbers of training data of source and target domains respectively. The height and width of the input are denoted as H and W. Due to the superior performance and training complexity, we follow the self-training framework Zou et al. (2018); Wang et al. (2021b); Zhang et al. (2021); Zheng & Yang (2021); Araslanov & Roth (2021) and adopt the mean teacher model Tarvainen & Valpola (2017); French et al. (2017); Araslanov & Roth (2021); Hoyer et al. (2022) to achieve an end-to-end self-training learning process, avoiding the cumbersome iterative training stage Zhang et al. (2021); Araslanov & Roth (2021).

Previous works Zou et al. (2018; 2019); Mei et al. (2020); Wang et al. (2021b); Zheng & Yang (2021); Araslanov & Roth (2021) mostly adopt a pre-trained model of the source domain as the initial model, and conduct an iterative process between pseudo-label generation and target domain training. Instead, mean teacher model jointly trains the images of the source and target domains end-to-end. Therefore, it is crucial to coordinate the source and target domains in the training process. Since no reliable pseudo-labels of the target domain are available at the beginning of the training, prematurely introducing target domain training brings label noise and prevents the model from learning discriminative semantic features. Conversely, introducing target domain training late can bias the model toward the source distribution and trap it in a local optimum.

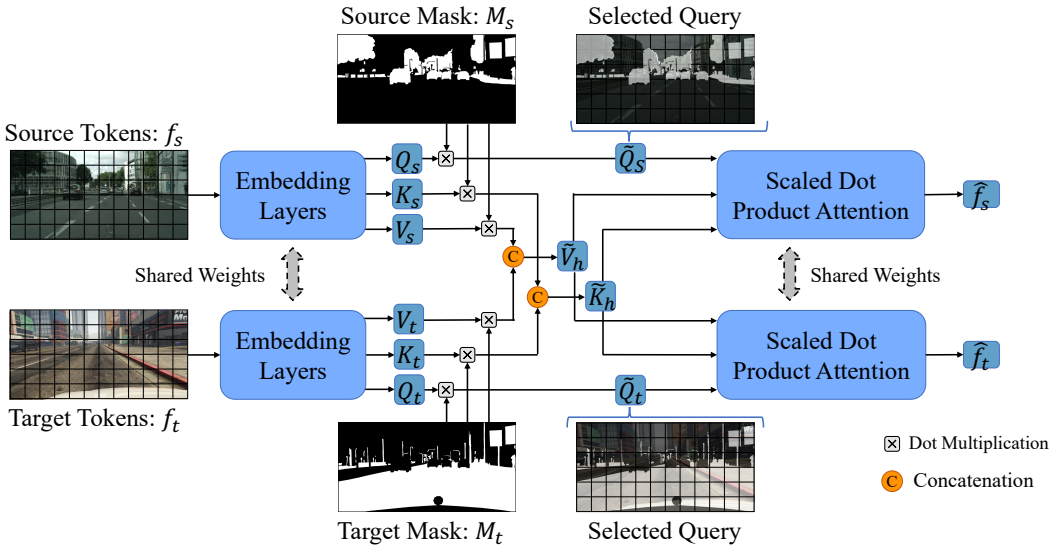


Figure 2: Illustration of the MHA module. Taking the feature tokens f_s, f_t from source and target domains as inputs, three embedding layers project these tokens to the corresponding query Q_i, key K_i, and value tokens V_i respectively, where $i \in \{s, t\}$. According to the feature grouping strategy, grouped key tokens (white region) from both domains forms the hybrid-domain key $\widetilde{K}_h = [\widetilde{K}_s; \widetilde{K}_t]$, as does the hybrid-domain query $\widetilde{V}_h = [\widetilde{V}_s; \widetilde{V}_t]$. The grouped source query $\widetilde{Q}_s$ and target query $\widetilde{Q}_t$ are multiplied respectively with hybrid-domain key $\widetilde{K}_h$ as the similarity matrix of the corresponding domain, respectively. Hybrid-domain value $\widetilde{V}_h$ is fused based on corresponding similarity matrix to generate hybrid-domain features $\hat{f}_s$ and $\hat{f}_t$ for the next MHA layer. The process of the complementary features (black region) can be derived in the same way.

To address the above issue, we propose a Source-Target Coordinated Training (STCT) framework to timely incorporate the target domain and the source domain during the training process. We argue that the coordination between source and target domain during training can be determined by the performance of the student model on the target domain. A well-performing student model means that the teacher model can provide reliable pseudo-labels. As a result, we propose to using the pseudo-accuracy of the target prediction $\boldsymbol{p}_t^i$ (output by the student model) on the target pseudo-label $\hat{\boldsymbol{y}}_t^i$ (output by the teacher model) as an anchor to control the participation of the target domain. The coordination weight is defined as:

$$Coor(\boldsymbol{p}_t^i, \hat{\boldsymbol{y}}_t^i) = Acc(\bar{\boldsymbol{y}}_t^i, \hat{\boldsymbol{y}}_t^i) \times (1 - e^{-iter \cdot \alpha}), \quad \text{where } \bar{\boldsymbol{y}}_t^{i,j} = \arg\max_k \boldsymbol{p}_t^{i,j,k} \tag{1}$$

Iteration step is denoted as $iter$ and α is a hyperparameter to control the ascent speed. Therefore, the final loss is formulated as $L = L_s + Coor(\boldsymbol{p}_t^i, \hat{\boldsymbol{y}}_t^i) \cdot L_t$. We adopt DAFormer Hoyer et al. (2022) as our teacher and student models. The STCT method is taken as the baseline for following experiments.

It is worth noting that a straightforward solution to solve the "when to align" is to choose a timing to involve the target domain with a fixed weight in the training process. Our method uses the accuracy of the student model on the target domain to dynamically determine the timing and the weight of engaging the target domain in training. Instead of adopting an exact threshold of accuracy to rigidly determine that the coordination weight is either 0 or 1, using a smoothing varying coordination weight can be taken as an advanced version and achieves better performance.

3.2 Multi-head hybrid attention module

Taking a pair of augmented source and target images, the student network in Fig. 1 first downsamples and reshapes the inputs (features) into a sequence of source tokens $f_s \in \mathbb{R}^{N \times d}$ and target tokens $f_t \in \mathbb{R}^{N \times d}$. The number of tokens in the i-th stage of Transformer architecture is denoted by $N = \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}$, and d is the number of channels. Then, these tokens are sent into Multi-head Self-Attention (MSA) layers to propagate and aggregate information.

In one MSA layer, token features $f \in \mathbb{R}^{N\times d}$ are projected into query features $Q \in \mathbb{R}^{N\times d_q}$, key features $K \in \mathbb{R}^{N\times d_k}$, and value features $V \in \mathbb{R}^{N\times d_v}$ by three embedding layers as demonstrated in Fig. 2. Then, value features V are aggregated together based on the similarities between query Q and key features K to form the token features $\hat{f}$ of the next MSA layer. The core self-attention function of the MSA layer is given as:

$$Attn_{self}(Q, K, V) = Softmax(\frac{QK^\top}{\sqrt{d_k}})V. \tag{2}$$

It is worth noting that aggregated value features V in the original MSA layer come from the same domain, *i.e.*, V is either from the source domain or the target domain.

To explicitly construct the interaction between the source and target features during feature learning process, we propose a Multi-head Hyrid-domain Attention (MHA) module to achieve feature alignment by fusing value features from both source and target domains. The MHA module first concatenates key features from two domains as hybrid-domain key $K_h = [K_s; K_t] \in \mathbb{R}^{2N\times d_k}$ and constructs hybrid-domain value $V_H = [V_s; V_t] \in \mathbb{R}^{2N\times d_v}$. Then, single-domain query (source query Q_s or target query Q_t) is multiplied with hybrid-domain key K_h to generate similarity matrix, which guides the linear weighted summation of hybrid-domain value V_h:

$$Attn_{hybrid}(Q_i, K_s, K_t, V_s, V_t) = Softmax(\frac{Q_i[K_s; K_t]^\top}{\sqrt{d_k}})[V_s; V_t], \text{ where } i \in \{s, t\}. \tag{3}$$

Since there are both self-attention ($Q_sK_s^\top$ and $Q_tK_t^\top$) for intra-domain feature fusion and cross-attention ($Q_sK_t^\top$ or $Q_tK_s^\top$) for inter-domain feature fusion, we call it the hybrid-attention module. Considering this fusion strategy takes all K and V features from both domains without grouping, we call it the "***non-grouping***" strategy. Primitive Transformer with MSA module only considers intra-domain information, since query, key, and value features all come from the same domain. Besides the intra-domain self-attention mechanism, our proposed MHA module also contains the inter-domain cross-attention mechanism to achieve feature fusion and alignment. However, directly using MHA leads to an unsatisfactory result. Under this circumstance, we conduct a series of experiments to analyse each component in MHA.

3.3 Deeper analysis on the MHA module

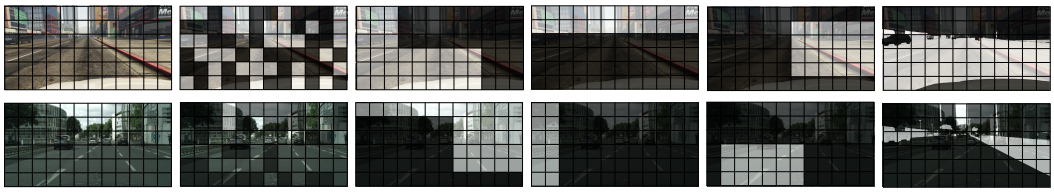
We provide our analysis from two aspects. The first one focuses on the feature grouping strategy to select suitable tokens for alignment in our MHA module. The second is to investigate the importance of source and target features in MHA. From the two aspects, we find some crucial strategies to achieve effective feature alignment. Following these strategies, we implement a bidirectional semantic-grouping MHA module that achieves outstanding performance.

3.3.1 Feature grouping strategies in the MHA

For the MHA module with the "non-grouping" strategy, we argue that source (target) query tokens Q_s (Q_t) tend to focus on the target (source) key tokens K_t (K_s) that are highly similar to them and easily aligned with them. These key tokens are easy samples for the corresponding query tokens. During training, these easy key tokens provide limited contribution for aligning source and target distributions, which leaves the alignment unsolved. In other words, building a shared common subspace for alignment is essential for domain adaptation Gopalan et al. (2011); Zhang et al. (2017); Ganin et al. (2016), and the easy key samples contribute little to pulling source/target distributions to the shared common subspace.

Therefore, we propose the feature grouping strategy to weaken the influence of these simple tokens by importing randomness. Specifically, we randomize query, key, and value tokens of source and target domains into two groups. The MHA module is applied independently in both groups to avoid the query tokens always focusing on the easily aligned key tokens of the other domain. The effectiveness of feature grouping strategy has been confirmed in literature Yu et al. (2022); Zeng et al. (2022) as well.

To validate our hypothesis, we design five grouping strategies from the perspectives of randomness, spatial continuity, and semantic integrity in feature tokens. For verifying the effect of randomness, the "***random-grouping***" strategy is designed to randomly divide the source and target domains,



(a) Without mask (b) Random mask (c) HVH mask (d) HVR mask (e) Cutout mask (f) Semantic mask

Figure 3: Illustration of (a) original feature tokens and (b)-(f) five masks of feature grouping strategies. Source and target tokens are listed separately in the first and second rows. Feature tokens in the white region are selected from Q_i, K_i, V_i and grouped as $\widetilde{Q}_i$, $\widetilde{K}_i$, $\widetilde{V}_i$, where $i \in \{s, t\}$. The complementary feature tokens in the black region constitute complementary groups $\widetilde{Q}_i^c$, $\widetilde{K}_i^c$, $\widetilde{V}_i^c$.

respectively, into two groups based on two random binary masks shown in Fig. 3(b). For spatial continuity, we present a Horizontal-Vertical Half grouping ("***HVH-grouping***") strategy to divide the features into two halves, horizontally or vertically. The HVH-grouping mask is shown in Fig. 3(c). To combine spatial continuity with randomness, we propose to randomly divide the image into two horizontal or vertical groups without equal division restriction, according to a Horizontal-Vertical Random (HVR) mask ("***HVR-grouping***" strategy) in Fig. 3(d). Besides, we also implement a random continuous division of the features based on the cutout mask, which is called the "***cutout-grouping***" strategy and shown in Fig. 3(e). For integrating semantic integrity into randomness and spatial continuity, we propose a "***semantic-grouping***" strategy, where the semantic mask illustrated in Fig. 3(f) selects the entire category region and divides the images into two groups based on class labels. The first four grouping strategies only require the size of the source and target images, while the latter semantic-grouping strategy also needs the ground-truth labels of the image. Considering the semantic categories from the source and target domains have an approximate spatial distribution and structure Tsai et al. (2018), we adopt a semantic mask from the source domain as the semantic mask of the unlabeled target domain to group features.

The goal of feature grouping strategies is to divide query, key, and value tokens into two exclusive groups where the MHA module is applied independently. We take the semantic-grouping strategy as an example. According to the semantic mask in Fig. 3(f), feature tokens Q_i, K_i, and V_i (where $i \in \{s, t\}$) of source and target domains are divided into the semantic group where $M_i = 1$ (white regions) and the complementary group where $M_i = 0$ (black regions). For the semantic group in white regions, key tokens $\widetilde{K}_s$ and $\widetilde{K}_t$ of source and target domains constitute hybrid-domain key tokens $\widetilde{K}_h = [\widetilde{K}_s; \widetilde{K}_t]$. Hybrid-domain value tokens $\widetilde{V}_h = [\widetilde{V}_s; \widetilde{V}_t]$ are composed in the same way. For the source (target) domain, hybrid-domain value tokens $\widetilde{V}_h$ is fused base on the similarity matrix, which is computed by the scaled dot-production between the single-domain query $\widetilde{Q}_s$ ($\widetilde{Q}_t$) and hybrid-domain key $\widetilde{K}_h$. The MHA module applied in the semantic group features is formulated as:

$$Attn_{hybrid}(\widetilde{Q}_i, \widetilde{K}_s, \widetilde{K}_t, \widetilde{V}_s, \widetilde{V}_t) = Softmax(\frac{\widetilde{Q}_i \cdot [\widetilde{K}_s; \widetilde{K}_t]^\top}{\sqrt{d}})[\widetilde{V}_s; \widetilde{V}_t], \quad \text{where } i \in \{s, t\}. \tag{4}$$

We emphasize that the source query $\widetilde{Q}_s$ and target query $\widetilde{Q}_t$ share the same hybrid-domain key $\widetilde{K}_h$ and value $\widetilde{V}_h$. Projecting the source and target domain spaces into a subspace with shared common basis $\widetilde{V}_h$ facilitates feature alignment. It is worth noting that only features in the white regions ($M_s = 1$ or $M_t = 1$) of the source and target domains are refined by Eq. 4. For the complementary group in black regions, the MHA module is applied as:

$$Attn_{hybrid}(\widetilde{Q}_i^c, \widetilde{K}_s^c, \widetilde{K}_t^c, \widetilde{V}_s^c, \widetilde{V}_t^c) = Softmax(\frac{\widetilde{Q}_i^c \cdot [\widetilde{K}_s^c; \widetilde{K}_t^c]^\top}{\sqrt{d}})[\widetilde{V}_s^c; \widetilde{V}_t^c], \quad \text{where } i \in \{s, t\}. \tag{5}$$

We only depict the process of Eq. 4 in the Fig. 2. The complementary part in Eq. 5 can be derived similarly.

[1]We use superscript c of a feature to denote the complementary feature.

3.3.2 IMPORTANCE OF SOURCE AND TARGET FEATURES IN MHA

As described in Sec. 3.2, the MHA module includes both intra-domain self-attention and inter-domain cross-attention mechanisms. The intra-domain self-attention comes from $\widetilde{Q}_s\widetilde{K}_s^\top$ and $\widetilde{Q}_t\widetilde{K}_t^\top$ in Eq. 4. The inter-domain cross-attention are from $\widetilde{Q}_s\widetilde{K}_t^\top$ and $\widetilde{Q}_t\widetilde{K}_s^\top$ in Eq. 4. We provide detailed studies on these four parts to investigate the importance of source and target features in MHA. Specifically, we name $\widetilde{Q}_s\widetilde{K}_s^\top$ and $\widetilde{Q}_t\widetilde{K}_t^\top$ as the intra-domain self-attention weight, while $\widetilde{Q}_s\widetilde{K}_t^\top$ and $\widetilde{Q}_t\widetilde{K}_s^\top$ as the inter-domain cross-attention weight.

The inter-domain cross-attention weight $\widetilde{Q}_s\widetilde{K}_t^\top$ guides the fusion of target value tokens $\widetilde{V}_t^\top$ to rebuild $\widetilde{Q}_s$, which tries to pull the source distribution to the target. So we consider it as a feature fusion on source-to-target direction. Similarly, $\widetilde{Q}_t\widetilde{K}_s^\top$ guides the fusion of source value tokens $\widetilde{V}_s^\top$ on target-to-source direction. Either of them is a ***unidirectional*** cross-attention mechanism. The ***bidirectional*** cross-attention mechanism includes both source-to-target and target-to-source directions, as described in Sec. 3.3.1. For a unidirectional cross-attention mechanism, take target-to-source direction as example, the Eq. 4 is changed to:

$$Attn_{hybrid}(\widetilde{Q}_s, \widetilde{K}_s, \widetilde{K}_t, \widetilde{V}_s, \widetilde{V}_t) = Softmax(\frac{\widetilde{Q}_s \cdot [\widetilde{K}_s; \widetilde{K}_t]^\top}{\sqrt{d}})[\widetilde{V}_s; \widetilde{V}_t], \tag{6}$$

$$Attn_{hybrid}(\widetilde{Q}_t, \widetilde{K}_t, \widetilde{V}_t) = Softmax(\frac{\widetilde{Q}_t \cdot (\widetilde{K}_t)^\top}{\sqrt{d}})\widetilde{V}_t. \tag{7}$$

Compared to the bidirectional cross-attention mechanism in Eq. 4, the cross-attention mechanism of source-to-target direction is only applied to the source domain as Eq. 6 and not to the target domain as Eq. 7. The unidirectional cross-attention mechanism of target-to-source direction is performed by exchanging the source and target domains. According to the experiments in Tab. 2, the unidirectional cross-attention mechanism does not bring significant performance improvement. We argue that the bidirectional cross-attention mechanism facilitates the projection of source and target features into a shared common subspace, minimizing the discrepancy in distribution between the projected source and target domains. Instead, the unidirectional cross-attention mechanism only attempts to project one domain (source or target) into the other domain, while keeping the other domain unchanged. There is also literature that shares the same conclusion on the shared common subspace in domain adaptation Gopalan et al. (2011); Zhang et al. (2017); Ganin et al. (2016).

Besides, we also conduct experiments to investigate the proportion between the intra-domain self-attention weight and the inter-domain self-attention weight during the feature fusion. We find that both weights are essential in the MHA module, and keeping them as an evenly matched ratio would achieve the best performance. The details of these experiments are presented in supplementary material due to the page limitation.

4 EXPERIMENTS

We conduct experiments on the two standard benchmark settings, namely, "GTAV Richter et al. (2016) to Cityscapes Cordts et al. (2016)" and "SYNTHIA Ros et al. (2016) to Cityscapes Cordts et al. (2016)", where GTAV Richter et al. (2016) and SYNTHIA Ros et al. (2016) are adopted as labeled source domain, and Cityscapes Cordts et al. (2016) is taken as unlabeled target domain to evaluate the adaptation performance. The details of our experimental settings are given in supplementary material.

4.1 ABLATION STUDIES FOR COORDINATION WEIGHT

As the basis of our framework, we first investigate the effects of the coordinate weight. Coordinate weight α in Eq. 1 is designed to achieve harmonious training between the source and target domains. We plot the curves of coordinate weights with different α on the left of Fig. 4. The corresponding performance is shown on the right of Fig.4. Setting $\alpha = 0$ means that the target domain is not involved and only the source domain is available for model training. As the α increases, the earlier the target domain is involved in the training. In addition, we also report the performance without the coordination weight, *i.e.*, the target domain is involved at the beginning of the training process with the same weight as the source domain. As shown in the dashed line, the model that introduces the target domain at the beginning only achieve 54.0, which is 2.63% higher than that of the model only

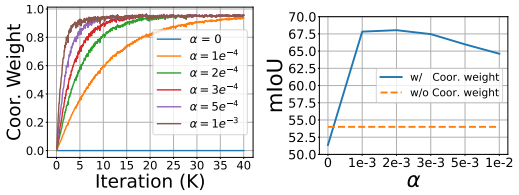


Figure 4: Coordination weight curves and performance comparison between various α. Setting $\alpha = 0$ indicates that target domain is not involved in the training. Our method reaches 54.0 without coordination weight as shown in the dashed line.

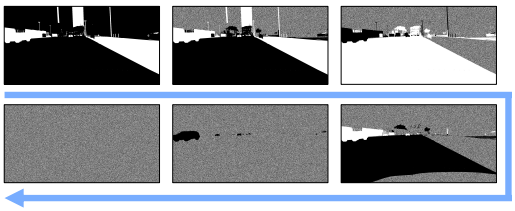


Figure 5: Illustration of progressively scattered semantic masks. The semantic masks with 0, 4, 8, 12, 14 (all) scattered classes are listed in the direction of the arrow.

trained on the source domain (51.37 at $\alpha = 0$). We take $\alpha = 2e^{-4}$ as our default setting according to the experimental results.

4.2 Ablation studies for the MHA module

Effects of the feature grouping strategies The feature grouping strategy determines which features are selected and fused in the MHA module. Five feature grouping strategies are designed according to the guidance of randomness, spatial continuity, and semantic integrity, as shown in Fig. 3. Performance in Tab. 2 indicates that all three factors are indispensable. First, randomly selecting and aggregating features from the source and target domain can provide more diversity. Replacing the single-domain features of the source and target domains with various hybrid-domain features enables the model to learn domain-invariant class representations under different feature distributions. Second, spatial continuity can maintain the contextual information, *i.e.*, features of the surrounding pixels, which is proven to enhance the feature representation on semantic segmentation Zhang et al. (2018); Yuan et al. (2020); Li et al. (2021); Jin et al. (2021). Moreover, spatial continuity preserves as many features as possible for minority categories, whose performance is the bottleneck of domain adaptation. Third, semantic integrity guarantees that the model can learn discriminative features for each class on the source domain, which is the basis for effective feature adaptation on the target domain. As a result, the semantic-grouping strategy achieves the best performance.

Along the semantic-grouping strategy, we conduct experiments to deeply verify the importance of semantic integrity. As illustrated in Fig. 3(f), semantic-grouping strategy divides feature tokens of one image into two exclusive parts, according to the semantic mask. For example, a source image in GTAV Richter et al. (2016) with 14 classes is divided into a black region with 7 classes and a white region with 7 classes, as shown in the left corner of Fig. 5. Each region contains all feature tokens of the corresponding 7 classes. To further verify the significance of semantic integrity in the MHA module, we scatter the features of the class present in the black part or the white part into two parts uniformly. This strategy is called the ***scattered-semantic (SS) grouping*** strategy. We gradually increase the number of scattered classes. The produced six masks arranged by arrows in Fig. 5 correspond to the cases where 0, 1, 4, 8, 12, and 14 categories are scattered. When all classes (14 classes in this image) are scattered, the class-scattered grouping strategy is the same as the random-grouping strategy in Fig. 3(b). The performance of six scattered-semantic grouping strategies is given on the right side of Tab. 2. As the number of scattered classes increases, the performance decreases from 68.05 (the same as the semantic-grouping strategy) to 63.37 (the same as the random-grouping strategy).

Effects of the bidirectional cross-attention mechanism We compare the bidirectional cross-attention mechanism with two unidirectional cross-attention mechanisms, and the results are listed in Tab. 2. It can be seen that the bidirectional cross-attention mechanism achieves much better performance than others. There is an interesting conclusion that the performance of only using the source-to-target cross-attention mechanism is even worse than no cross-attention baseline, which may be due to the low confidence of the pseudo labels in the target domain.

MHA in different Transformer stages As shown in Tab. 1, since the MHA numbers in four stages go up and then down, the adaptation performance of our method first arises from 61.26 to 66.93, then drops back to 61.30. We believe the core factor that affects performance is the number of blocks in each stage. More MHA modules are applied, the more it helps the model learn domain-invariant

Table 1: **Ablation study of the MHA module in each Transformer stage.** The first line shows the performance without the MHA module.

Stage1	Stage2	Stage3	Stage4	mIoU
–	–	–	–	56.65
✓	–	–	–	57.21
–	✓	–	–	58.87
–	–	✓	–	65.70
–	–	–	✓	57.51
✓	✓	–	–	60.71
–	✓	✓	–	66.34
–	–	✓	✓	66.79
✓	✓	✓	–	66.29
–	✓	✓	✓	67.87
✓	✓	✓	✓	**68.05**

Table 2: **Ablation study of feature grouping strategies**. The SSG strategy indicates the scattered-semantic grouping strategy. Source-to-target and target-to-source directions are denoted by S2T and T2S respectively.

Strategy	mIoU	SSG Strategy	mIoU
non-grouping	61.91	non classes	**68.05**
random-grouping	63.37	1 classes	67.65
HVH-grouping	64.55	4 classes	65.94
HVR-grouping	64.93	8 classes	65.83
cutout-grouping	65.07	12 classes	64.58
semantic-grouping	**68.05**	all (14) classes	63.37

Baseline without cross-attention mechanism	56.65
Unidirectional cross-attention mechanism (S2T)	55.91
Unidirectional cross-attention mechanism (T2S)	58.38
Bidirectional cross-attention mechanism	**68.05**

Table 3: Comparison with state-of-the-arts on GTAV/Synthia to Cityscapes benchmarks. The **first** and second highest scores are represented by bold font and underline respectively.

	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU
GTA5 → Cityscapes																				
ADVENT Vu et al. (2019)	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
CBST Zou et al. (2018)	91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	42.8	45.9
DACS Tranheden et al. (2021)	89.9	39.7	87.9	30.7	39.5	38.5	46.4	52.8	88.0	44.0	88.8	67.2	35.8	84.5	45.7	50.2	0.0	27.3	34.0	52.1
DPL-Dual Cheng et al. (2021)	92.8	54.4	86.2	41.6	32.7	36.4	49.0	34.0	85.8	41.3	86.0	63.2	34.2	87.2	39.3	44.5	18.7	42.6	43.1	53.3
SAC Araslanov & Roth (2021)	90.4	53.9	86.6	42.4	27.3	45.1	48.5	42.7	87.4	40.1	86.1	67.5	29.7	88.5	49.1	54.6	9.8	26.6	45.3	53.8
CorDA Wang et al. (2021a)	94.7	63.1	87.6	30.7	40.6	40.2	47.8	51.6	87.6	47.0	89.7	66.7	35.9	90.2	48.9	57.5	0.0	39.8	56.0	56.6
ProDA Zhang et al. (2021)	87.8	56.0	79.7	46.3	44.8	45.6	53.5	53.5	88.6	45.2	82.1	70.7	39.2	88.8	45.5	59.4	1.0	48.9	56.4	57.5
DAFormer Hoyer et al. (2022)	95.7	70.2	89.4	53.5	48.1	**49.6**	55.8	**59.4**	**89.9**	47.9	**92.5**	**72.2**	**44.7**	**92.3**	**74.5**	78.2	65.1	55.9	**61.8**	**68.3**
Ours	**96.2**	**73.1**	**89.6**	**59.4**	**48.5**	43.8	**57.6**	53.5	89.7	**49.1**	91.8	69.9	41.9	91.9	68.8	**80.0**	**69.3**	**57.9**	61.1	68.1
Synthia → Cityscapes																				
ADVENT Vu et al. (2019)	85.6	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4	–	84.1	57.9	23.8	73.3	–	36.4	–	14.2	33.0	41.2
CBST Zou et al. (2018)	68.0	29.9	76.3	10.8	1.4	33.9	22.8	29.5	77.6	–	78.3	60.6	28.3	81.6	–	23.5	–	18.8	39.8	42.6
DACS Tranheden et al. (2021)	80.6	25.1	81.9	21.5	2.9	37.2	22.7	24.0	83.7	–	90.8	67.6	38.3	82.9	–	38.9	–	28.5	47.6	48.3
DPL-Dual Cheng et al. (2021)	83.5	38.2	80.4	1.3	1.1	29.1	20.2	32.7	81.8	–	83.6	55.9	20.3	79.4	–	26.6	–	7.4	46.2	43.0
SAC Araslanov & Roth (2021)	89.3	47.2	85.5	26.5	1.3	43.0	45.5	32.0	87.1	–	89.3	63.6	25.4	86.9	–	35.6	–	30.4	53.0	52.6
CorDA Wang et al. (2021a)	**93.3**	**61.6**	85.3	19.6	5.1	37.8	36.6	42.8	84.9	–	90.4	69.7	41.8	85.6	–	38.4	–	32.6	53.9	55.0
ProDA Zhang et al. (2021)	87.8	45.7	84.6	37.1	0.6	44.0	54.6	37.0	**88.1**	–	84.4	**74.2**	24.3	**88.2**	–	51.1	–	40.5	45.6	55.5
DAFormer Hoyer et al. (2022)	84.5	40.7	**88.4**	41.5	6.5	**50.0**	**55.0**	**54.6**	86.0	–	89.8	73.2	48.2	87.2	–	**53.2**	–	53.9	**61.7**	**60.9**
Ours	81.8	40.0	**88.4**	**44.7**	**8.1**	47.2	49.2	49.0	86.5	–	**92.2**	73.8	**49.6**	86.3	–	42.1	–	**56.7**	61.0	59.8

features. When the MHA module is applied incrementally in four stages, consistent performance improvement is achieved and up to 68.05.

4.3 Comparison to state-of-the-art methods

To show the superiority of the proposed method, we report adaptation performance in terms of mIoU (%) on two benchmarks in Fig. 3. Our method achieves 68.05 and 59.8 performance on GTAV-to-Cityscapes and Synthia-to-Cityscapes benchmarks, respectively. Compared to the state-of-the-art method DAFormer Hoyer et al. (2022), we achieve a comparable performance without using DACS Tranheden et al. (2021) augmentation. Considering the source pre-trained model can provide a better initialization model, we train our STCT framework with a source pre-trained model, which achieve a new SOTA performance **69.18** in mIoU on GTAV-to-Cityscapes benchmark. To further verify the effectiveness of coordination weight, experiments of STCT framework based on the source pre-trained model but without coordination weight is conducted and only achieves 67.06 in mIoU on GTAV-to-Cityscapes benchmarks, which is inferior to the counterpart with the coordination weight.

5 Conclusion

In this work, we presented a Source-Target Coordinated Training Framewor (STCT) framework based on the coordinate weight for unsupervised domain adaptation on semantic segmentation. The proposed STCT framework solves the problem of coordination between the learning of discriminative category features from source domain and the learning of feature distribution of target domain. Moreover, to explicitly construct the interaction and bridge the gap between different domains, we propose a Multi-head Hybrid-Attention (MHA) module. The MHA module consists of intra-domain self-attention and inter-domain cross-attention to achieve feature fusion and alignment. Exhaustive experiments are conducted to study the effects of various factors in the MHA module.

REFERENCES

Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15384–15394, 2021.

Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 357–366, 2021.

Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, Fang Wen, and Wenqiang Zhang. Dual path learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9082–9091, 2021.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. *Int. Conf. Learn. Representations.*, 2017.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6639–6648, 2019.

Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *2011 international conference on computer vision*, pp. 999–1006. IEEE, 2011.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *arXiv e-prints*, pp. arXiv–2012, 2020.

Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Zhenchao Jin, Bin Liu, Qi Chu, and Nenghai Yu. Isnet: Integrate image-level and semantic-level context for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7189–7198, 2021.

Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12975–12984, 2020.

Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6936–6945, 2019.

Zechao Li, Yanpeng Sun, Liyan Zhang, and Jinhui Tang. Ctnet: Context-based tandem network for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Yahao Liu, Jinhong Deng, Xinchen Gao, Wen Li, and Lixin Duan. Bapa-net: Boundary adaptation and prototype alignment for cross-domain semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8801–8811, 2021a.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021b.

Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *European conference on computer vision*, pp. 415–430. Springer, 2020.

Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pp. 102–118. Springer, 2016.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3234–3243, 2016.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1379–1389, 2021.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7472–7481, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2517–2526, 2019.

Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8515–8525, 2021a.

Yuxi Wang, Junran Peng, and ZhaoXiang Zhang. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9092–9101, 2021b.

Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*, 2021.

Jinyu Yang, Weizhi An, Sheng Wang, Xinliang Zhu, Chaochao Yan, and Junzhou Huang. Label-driven reconstruction for domain adaptation in semantic segmentation. In *European conference on computer vision*, pp. 480–498. Springer, 2020.

Tan Yu, Gangming Zhao, Ping Li, and Yizhou Yu. Boat: Bilateral local attention vision transformer. *arXiv preprint arXiv:2201.13027*, 2022.

Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *European conference on computer vision*, pp. 173–190. Springer, 2020.

Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Ouyang Wanli, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. *arXiv preprint arXiv:2204.08680*, 2022.

Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 7151–7160, 2018.

Jing Zhang, Wanqing Li, and Philip Ogunbona. Joint geometrical and statistical alignment for visual domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1859–1867, 2017.

Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12414–12424, 2021.

Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, 129(4):1106–1120, 2021.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 289–305, 2018.

Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5982–5991, 2019.