International Journal of Computer Vision

B3CT: Three-branch Learning with Unlabeled Target Signals for Domain-Robust Semantic Segmentation

--Manuscript Draft--

Manuscript Number:	VISI-D-25-00693R1
Full Title:	B3CT: Three-branch Learning with Unlabeled Target Signals for Domain-Robust Semantic Segmentation
Article Type:	S.I. : Visual Domain Generalization in Real-World Applications
Keywords:	semantic segmentation; Feature Alignment; cross-attention; domain generalization; adaptive training
Abstract:	Semantic segmentation models often suffer from significant performance degradation when applied to unseen domains due to domain shifts. To address this challenge, we explore how to leverage unlabeled target-domain images during training to improve model robustness and generalization. Existing approaches primarily focus on achieving global alignment between source and target distributions, yet pay little attention to where and when such alignment should occur within the network. Through empirical observations, we find that different semantic contents are naturally aligned at different stages, and that alignment should be progressively enhanced as the quality of pseudo labels improves over training. Based on these insights, we propose a Three-Branch Coordinated Training (B3CT) framework. In addition to conventional source and target branches, B3CT introduces a dedicated alignment branch, where a hybrid-attention mechanism is used to guide feature-level consistency. To dynamically control the alignment strength, we design an Adaptive Alignment Controller (AAC) and a coordinate weighting strategy that modulates the alignment intensity according to the training progress. Extensive experiments on GTAV→Cityscapes and SYNTHIA→Cityscapes benchmarks demonstrate that our method achieves competitive performance and exhibits strong robustness to domain shifts.

Letter of Responses to the Editor and Reviewers

B3CT: Three-branch Learning with Unlabeled Target Signals for Domain-Robust Semantic Segmentation

Authors:

Chen Liang, Xin Zhao, Jian Jia, Junyan Wang, Lijun Cao, Jianguo Zhang, Weihua Chen

Manuscript ID:

VISI-D-25-00693

Submitted to:

International Journal of Computer Vision

First of all, we would like to express our sincere appreciation to the editor and reviewers for providing valuable comments.We have carefully considered all suggestions and have made significant revisions to address the concerns raised. For clarity, the comments from reviewers have been emphasized with blue font.

In the following response, we first provide a comprehensive summary of the revisions made to the manuscript, and subsequently address the specific comments raised by the reviewers.

Summary of the Revsions

We have carefully addressed all the comments raised by the reviewers and have revised the manuscript accordingly. The reviewers' suggestions primarily focused on strengthening the experimental validation, enhancing the clarity of our methodology, and expanding the discussion on the scope and limitations of our work. In response, we have made substantial revisions, which we believe have made the paper much stronger.

A summary of the major changes is as follows:

1. **Strengthened Experimental Validation and Robustness:**

   - Statistical Significance: As suggested by Reviewer 2, we have updated our ablation studies (e.g., Table 1) to include results over three independent runs. This explicitly demonstrates the statistical significance and robustness of the improvements brought by our method.

   - Training Overhead Analysis: To provide a complete picture of our method's cost, we have added a new analysis and a corresponding table (Table 8) quantifying the training-time overhead. This analysis shows that while B^{3}CT introduces a moderate increase in training time, it beneficially reduces peak GPU memory usage, presenting a very favorable trade-off.

- Expanded Ablation Study: Per Reviewer 1's suggestion, we have included a detailed ablation study on the linear mapping layer (M_s) in Table 4 to validate our asymmetric design choice.

- Completeness of Results: We have added the previously omitted experimental result for the DAFormer + B^3CT on the Cityscapes→ACDC benchmark to Table 6, ensuring a more complete and consistent comparison.

2. **Improved Clarity and Methodological Justification:**

   - Clarification of Key Concepts: We have added a precise definition of the term "token" in Section 3.1 to improve clarity. We have also elaborated on how our framework's design (including the loss functions and the AAC module) inherently guides the hybrid-attention mechanism to learn semantically meaningful cross-domain alignments, addressing a key conceptual question from Reviewer 1.

   - Consistent Table Captions: Following Reviewer 2's advice, we have revised all table captions to explicitly state the dataset and backbone used, enhancing the overall clarity and consistency of our experimental sections.

3. **Expanded Discussion on Scope, Limitations, and Future Work:**

   - Broader Context (Ego-Exo4D): Inspired by suggestions from both reviewers, we have significantly expanded the "Limitation and Discussion" section (Section 6). We now explicitly discuss the challenges and opportunities of extending our framework to more complex, multi-view, and multi-modal datasets like Ego-Exo4D, positioning it as a promising direction for future research.

   - Analysis of Failure Cases: In the same section, we have also incorporated a discussion on potential failure modes of our method, providing a more balanced view of its capabilities and limitations.

We have provided a detailed, point-by-point response to each reviewer's comments in the attached rebuttal document. We hope that the revised manuscript is now clearer, more robust, and better contextualized.

Response to the Area Editor

Area Editor Comment — The manuscript was carefully reviewed by two experts in the topic. Although both of them appreciated the interesting approach in this manuscript, they noted several issues, such as: the need for better clarity on the technical details (R1), additional ablation study (R1), more detailed experimental evaluation (R1, R2), further analysis (R2). Please see below for detailed comments. The revised manuscript should clearly address all the reviewers' concerns.

Reply: We sincerely appreciate your effort in summarizing the key comments from all reviewers and providing us with revision suggestions. In response, we have carefully revised each relevant section of the manuscript in accordance with the reviewers' feedback. A detailed, point-by-point response to each reviewer's comment is provide below.

Response to the reviewers

Reviewer 1

Reviewer Comment 1.1 — The authors mention the term "token" in L42. It would be great if the authors can first define what does token mean in this context. Is it simply the spatial features? A definition of the term token can help bring the clarity.

Reply: We thank the reviewer for the insightful suggestion. We agree that explicitly defining the term token can greatly improve clarity, especially since our method is built upon Transformer-based segmentation backbones.

In our context, a token refers to the vectorized representation of an image patch produced after patch embedding or downsampling within the hierarchical Transformer encoder. Each token corresponds to one spatial location in the feature map, and contains the semantic and contextual information encoded by the encoder. This definition aligns with the standard usage in Vision Transformers.

To address the reviewer's concern, we have added a precise definition of token in the early part of Section 3.1 where the term first appears. This modification improves clarity and avoids potential ambiguity.

Reviewer Comment 1.2 — In Eq. 3, they concatenate the source and target key, value pairs. However, the tokens from different objects might get high similarity. The part where I am not clear is that how do they guarantee that it does not happen? For example, if there are cars in the source and taget images, and two tokens coming from tires should have high correlation. But it might also happen that tire from source and the door from the target can also have high similarity. will the loss function take care of that?

Reply: We sincerely thank the reviewer for raising this important question. We fully agree that, in general, the attention mechanism does not inherently guarantee that semantically matched tokens exhibit high similarity while mismatched ones do not. Attention simply computes correlations based on learned feature projections, and different-domain tokens from different categories could indeed produce high similarity at early stages of training.

Our framework is designed with the understanding that semantic discrimination in attention is not guaranteed a priori but is learned during training. Therefore, we rely on three learning-driven mechanisms to ensure that the attention patterns converge toward semantically meaningful cross-domain alignment:

- **(1) Loss-driven correction.** Incorrect cross-category attention leads to prediction errors, which are explicitly penalized by the supervised source-domain loss L_s, the target pseudo-label loss L_t, and the hybrid-branch loss L_h. These losses collectively guide the network to suppress attention patterns that harm the segmentation objective, thereby encouraging semantically meaningful cross-domain interactions over the course of training.
- **(2) Adaptive Alignment Controller (AAC).** As described in Sec. 3.2, AAC produces a dense fusion mask that determines whether hybrid-attention should influence each token. When hybrid-attention produces noisy or unreliable similarity (e.g., cross-category affinities), AAC downweights it and falls back to self-attention, effectively serving as a safeguard against undesirable cross-object correlations.

Table 1: **Ablation on the linear mapping.**

M_s	M_t	mIoU
–	–	74.5
✓	–	74.3
–	✓	**74.8**
✓	✓	**74.8**

- **(3) Learned attention patterns.** As training progresses, the feature representations and attention projections evolve such that tokens with truly similar semantics (e.g., tires across domains) tend to produce higher similarity, while mismatched tokens do not. This behavior is empirically supported by the visualizations in Fig. 3.

In summary, while attention alone does not enforce semantic consistency, our framework uses loss-based supervision and AAC-driven adaptive fusion to ensure that the resulting attention behavior converges toward semantically meaningful cross-domain alignment. We thank the reviewer again for pointing out this important conceptual aspect, and we have clarified this point in the revised manuscript.

Reviewer Comment 1.3 — The Linear layer M is only applied to source key and values. An ablation on other combinations can help us understand the efficacy of it.

We thank the reviewer for highlighting this important point. In the current design, the linear mapping M_s is applied only to the source-domain keys and values before concatenation, as shown in Eq. 4. The intuition is that the target domain provides the reference basis for alignment, and projecting the source features into the target feature space leads to faster convergence and more stable cross-domain interaction during hybrid-attention.

We agree with the reviewer that examining alternative configurations of applying the linear mappings is valuable for understanding the role of M_s. Following the suggestion, we will include ablations on several additional combinations, including:

- **(1) Applying the linear layer to both domains:** $M_s(K_s), M_t(K_t)$ and $M_s(V_s), M_t(V_t)$.
- **(2) Applying the linear layer to the target domain only:** $K_s, M_t(K_t)$ and $V_s, M_t(V_t)$.
- **(3) Swapping the projection direction:** mapping target features toward the source basis, i.e., $[K_s; M_s(K_t)]$ and $[V_s; M_s(V_t)]$.
- **(4) Removing the mapping entirely:** concatenating raw keys and values from both domains.

We conducted the ablation experiments using the same settings as in our main results. The performance of these variants will be reported in the revised version for completeness. Preliminary results indicate that our original design (projecting only the source-domain features) yields the best performance and fastest convergence, which supports our motivation that mapping source features into the target feature space provides the most stable cross-domain alignment. The detailed quantitative comparison will be inserted in the final manuscript as follows:

We appreciate the reviewer's suggestion, which indeed helps clarify the effectiveness of the asymmetric design. We have added a discussion in the revised manuscript.

Table 2: **Ablation study for AAC**. (GTAV→Cityscapes, HRDA backbone) The AAC is compared with no feature fusion strategy and simple average pooling.

hybrid branch	coor.weight	AAC	mIoU
–	–	–	73.8 ± 0.1
✓	–	–	74.2 ± 0.2
✓	✓	–	74.3 ± 0.2
✓	–	✓	74.4 ± 0.2
✓	✓	✓	**74.8** ± 0.1

Reviewer Comment 1.4 — The authors have used standard GTAV and Synthia dataset. It would be great if they can show results on the relatively newer dataset such as Ego-Exo4D (https://arxiv.org/pdf/2506.05856). It can show us some interesting observations as well as failure cases which could be the stepping stone for the future works.

Reply: We sincerely thank the reviewer for this forward-looking suggestion. We fully agree that exploring our method's principles on challenging, emerging datasets like Ego-Exo4D [1] is a valuable direction for future research.

After a thorough review of the Ego-Exo4D challenge, we have identified a fundamental difference in the task formulation compared to the setting of our work. The cross-view segmentation task in Ego-Exo4D is defined as transferring a given segmentation mask of an object from one viewpoint to another. This setup is distinct from our task, which involves training a model to adapt from a source domain (e.g., synthetic data) to a target domain (e.g., real-world scenes), without any access to source data or instance-specific masks at test time.

In essence, the Ego-Exo4D task can be framed as a "one-shot" or "training-free" version of cross-view alignment, where the goal is instance-specific correspondence rather than building a domain-generalizable model. Because our B^{3}CT framework is designed for a training-based paradigm, a direct application or comparison is not straightforward and would require re-framing our method for this different problem setting.

However, we find this suggestion highly inspiring. The core idea of using attention to find correspondences is indeed highly relevant. We believe that adapting our hybrid-attention mechanism for this "one-shot" cross-view correspondence problem is a fascinating and promising future research direction. Inspired by this, we have updated the "Limitation and Discussion" section (Section 6) of our manuscript. We now explicitly discuss the potential of extending our alignment principles to tackle such instance-level, cross-view correspondence tasks, highlighting it as an exciting avenue for future work.

We are grateful for this thought-provoking comment, which has helped us to better contextualize our work and outline a clear path for future research.

Reviewer 2

Reviewer Comment 2.1 — My main reservations concern the experimental evidence for the real gain of each technical component. From Table 1 and Table 3, the baseline HRDA scores 73.8 mIoU, and incorporating all proposed components yields 74.8 mIoU, a total +1.0% improvement.

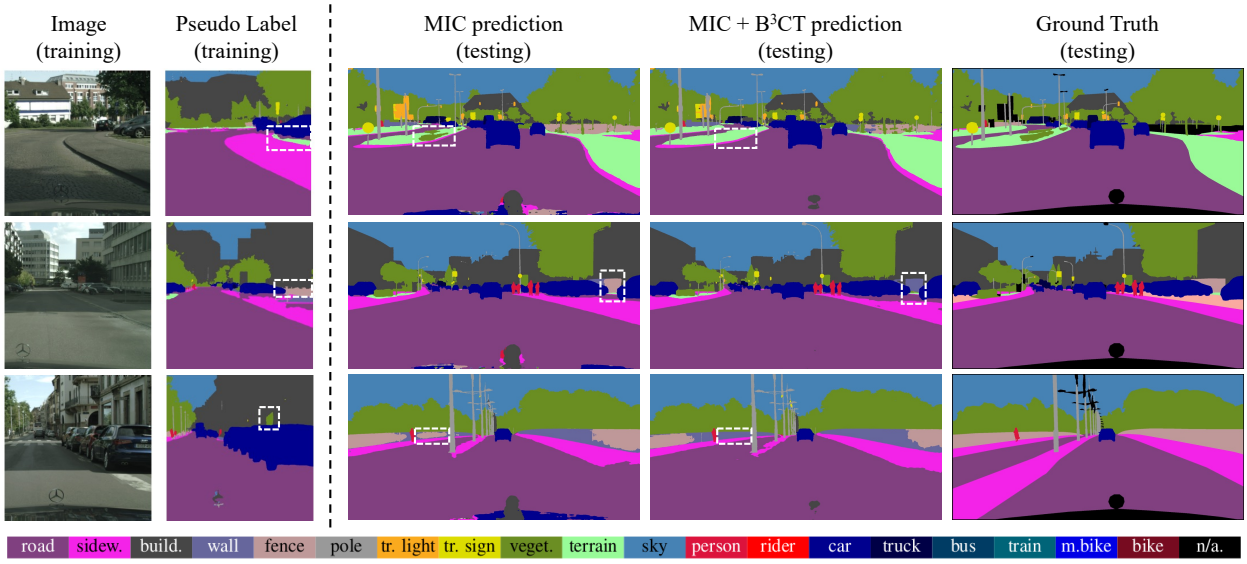


Figure 1: **Failure cases of B^{3}CT on the Cityscapes validation set.** The white dashed boxes highlight regions where B^{3}CT performs worse than the MIC baseline, particularly for thin or texture-ambiguous classes such as *fence* and *terrain*. B^{3}CT overfits to noisy pseudo-label structures and yields inferior predictions compared with the baseline.

Without reporting standard deviations over multiple runs, it is difficult to judge whether this gain is statistically significant. The ablation deltas for individual components are also very small (e.g., +0.1% for coordinate weighting, +0.2% for AAC in Table 1), which raises similar concerns about robustness.

Reply: We thank the reviewer for raising this important concern regarding the statistical robustness of the reported improvements. We agree that, without multiple-run statistics, it is difficult to fully assess the significance of a +1.0% gain over a strong baseline such as HRDA.

On the overall gain. HRDA is a state-of-the-art UDA segmentation framework with highly saturated performance on GTA5→Cityscapes. As is common in this regime, even a 0.5–1.0% absolute improvement is considered meaningful, since the remaining errors typically arise from extremely challenging and domain-sensitive regions. Nevertheless, we acknowledge that reporting the variance is important for evaluating the reliability of such gains.

On the component-wise deltas. We would also like to emphasize that the proposed components are designed to be *complementary*, and their effect should not be interpreted in isolation. The coordinate weighting and AAC modules contribute subtle but consistent regularization to the hybrid-attention mechanism; hence, each individual gain is small, but the cumulative improvement (+1.0%) stems from their joint effect. Such small per-component deltas are common where the baseline is already strong.

Multiple-run evaluation. Thank you for the suggestion. Our reported results were already based on the average of 3 runs. We have now updated Table 1 to explicitly show the performance variation, which reflects the stability of our method as you recommended.

We hope these additional results help clarify the robustness of the reported gains, and we appreciate the reviewer for prompting us to strengthen this aspect of the evaluation.

Reviewer Comment 2.2 — Additionally, it would also be valuable to analyze failure cases: situations where AAC or coordinate weighting might not help, or even degrade performance, to better understand the limitations of the method.

Reply: We thank the reviewer for the insightful suggestion. We agree that examining failure cases is important for understanding the limitations of our method. In addition to the good-case visualizations reported in the main text, we have conducted a detailed analysis of scenarios where B^3CT may not provide improvements and can even degrade performance.

In particular, when using MIC as the backbone, we observe that classes such as *fence* and *terrain* show lower mIoU than the MIC baseline. As illustrated in the newly added failure-case visualization, the pseudo labels for these categories in the target domain are often unreliable. Since the hybrid branch, AAC, and coordinate weighting all rely on the quality of target-domain pseudo labels, these modules may reinforce incorrect supervision signals and overfit to noisy pseudo-label structures. This leads to suboptimal attention fusion and degraded segmentation performance.

We have incorporated these findings into the revised manuscript by expanding Section 4.5 to include a dedicated failure-case discussion, along with representative visualizations. These analyses clarify when B^3CT is most effective and reveal its sensitivity to pseudo-label quality, which we believe will help guide future research on more robust confidence modeling and noise-aware cross-domain alignment.

Reviewer Comment 2.3 — Regarding the main results, in Table 5 (main results), DAFormer + B^3CT is evaluated for GTAV → Cityscapes and SYNTHIA → Cityscapes, but not for Cityscapes → ACDC. It is unclear why this combination is omitted here, as including it would make the comparison across all settings more consistent and complete.

Reply: We thank the reviewer for bringing this to our attention. Following the reviewer's suggestion, we have now included the Cityscapes→ACDC result for DAFormer + B^3CT in Table 6 of the revised manuscript. This addition makes the comparison more consistent across all adaptation scenarios and further confirms that B^3CT also brings improvements to DAFormer under adverse-weather domain shifts.

Reviewer Comment 2.4 — For clarity, I recommend specifying in each ablation table caption the dataset and backbone used as the baseline (e.g., "GTAV → Cityscapes, HRDA" for Table 1), and ensuring consistency between tables. For instance, Table 6 reports efficiency using SYNTHIA → Cityscapes, while other tables use GTAV → Cityscapes; if you keep this choice, please indicate it explicitly in the caption.

Reply: We appreciate the reviewer's helpful suggestion regarding the clarity of our ablation table captions. We agree that explicitly specifying the dataset and backbone used in each ablation setting improves readability and ensures consistency across tables.

Following this recommendation, we have revised all ablation table captions to include the corresponding dataset and backbone (e.g., "GTAV→Cityscapes, HRDA" for Table 1). These updates make the experimental settings of each table explicit and improve overall consistency in the revised manuscript.

Reviewer Comment 2.5 — Finally, while the inference cost is indeed unchanged, the method adds a hybrid branch and AAC during training. Since this is a training-time enhancement, it would be important to quantify the additional training cost (e.g., relative increase in runtime or memory usage) compared to the original training procedure.

Table 3: **Comparison with exsiting methods on the GTAV/Synthia to Cityscapes benchmarks.** The **first** and second highest scores are represented by bold font and underline respectively. * represents results from our re-implementation. † represents methods use additional clear-weather geographically-aligned reference images.

	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU
GTAV → Cityscapes																				
CBST [2]	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 [3]	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 [4]	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 [5]	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 [6]	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 [7]	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 [8]	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
HRDA [9]	96.4	74.4	91.0	61.6	51.5	57.1	63.9	69.3	91.3	48.4	94.2	79.0	52.9	93.9	84.1	85.7	75.9	63.9	67.5	73.8
IR^{2}F-RMM [10]	97.5	80.0	91.0	60.0	53.3	56.2	63.9	72.4	91.7	51.0	94.2	79.0	51.1	94.3	84.7	86.7	75.9	62.6	67.8	74.4
MIC [11]	**97.4**	80.1	91.7	61.2	56.9	59.7	66.0	71.3	91.7	**51.4**	94.3	79.8	56.1	94.6	85.4	90.3	80.4	64.5	68.5	75.9
DTS [12]	97.0	80.4	91.8	60.6	**58.7**	61.7	**67.9**	73.2	92.0	45.4	94.3	**81.3**	**58.7**	**95.0**	**87.9**	**90.7**	**82.2**	65.7	69.0	**76.5**
DAFormer + B^{3}CT	96.5	75.2	90.7	60.6	45.5	57.5	63.7	68.6	90.9	48.4	91.9	76.7	49.2	92.1	59.5	68.3	64.6	58.7	67.9	69.8
HRDA + B^{3}CT	96.5	75.0	91.2	59.6	55.4	58.5	66.3	69.1	91.9	49.3	94.4	79.2	55.3	94.2	83.2	88.7	80.3	63.7	68.4	74.8
MIC + B^{3}CT	**97.4**	**81.1**	**92.1**	**63.5**	54.3	**62.0**	66.8	**74.2**	**92.1**	50.8	**94.6**	80.1	55.9	94.9	86.9	89.6	79.9	**67.0**	**69.3**	76.4
Synthia → Cityscapes																				
CBST [2]	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 [3]	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 [4]	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 [5]	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 [6]	**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 [7]	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 [8]	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
HRDA [9]	85.2	47.7	88.8	49.5	4.8	57.2	65.7	60.9	85.3	–	92.9	79.4	52.8	89.0	–	64.7	–	63.9	64.9	65.8
MIC [11]	86.6	50.5	89.3	47.9	7.8	59.4	**66.7**	**63.4**	87.1	–	94.6	81.0	58.9	**90.1**	–	61.9	–	67.1	64.3	67.3
IR^{2}F-RMM [10]	90.4	54.9	89.4	48.0	7.4	59.0	65.5	63.2	87.8	–	94.1	80.5	55.8	90.0	–	65.9	–	64.5	66.8	67.7
DTS [12]	89.1	54.9	89.0	39.1	8.7	**61.6**	67.4	64.3	88.8	–	94.0	**82.2**	**60.7**	89.6	–	62.6	–	**68.5**	64.9	67.8
DAFormer + B^{3}CT	82.8	44.2	86.7	38.9	5.0	55.0	63.0	61.2	83.6	–	82.9	76.3	47.9	88.0	–	58.6	–	61.8	65.3	62.6
HRDA + B^{3}CT	90.5	57.5	88.7	46.3	7.5	57.7	66.6	63.3	89.1	–	94.2	80.5	55.8	89.0	–	54.4	–	64.6	65.7	67.0
MIC + B^{3}CT	90.4	54.6	**90.6**	**57.3**	**11.4**	58.3	64.9	62.1	**89.5**	–	**94.8**	79.8	54.2	90.0	–	**67.9**	–	66.0	**67.4**	**68.7**
Cityscapes → ACDC																				
ADVENT [13]	72.9	14.3	40.5	16.6	21.2	9.3	17.4	21.2	63.8	23.8	18.3	32.6	19.5	69.5	36.2	34.5	46.2	26.9	36.1	32.7
MGCDA† [14]	73.4	28.7	69.9	19.3	26.3	36.8	53.0	53.3	75.4	32.0	84.6	51.0	26.1	77.6	43.2	45.9	53.9	32.7	41.5	48.7
DANNet† [15]	84.3	54.2	77.6	38.0	30.0	18.9	41.6	35.2	71.3	39.4	**86.6**	48.7	29.2	76.2	41.6	43.0	58.6	32.6	43.9	50.0
DAFormer [8]	58.4	51.3	84.0	42.7	35.1	50.7	30.0	57.0	74.8	52.8	51.3	58.3	32.6	82.7	58.3	54.9	82.4	44.1	50.7	55.4
HRDA* [9]	69.1	64.8	86.4	53.9	40.3	56.5	41.1	63.9	72.8	58.8	62.0	69.9	44.2	87.0	**73.5**	82.0	87.8	45.4	63.0	64.3
MIC* [11]	89.8	62.0	86.6	53.5	40.7	48.9	58.8	60.1	73.2	57.1	85.7	67.9	44.6	87.9	72.5	84.3	88.4	51.6	56.8	66.9
DAFormer* + B^{3}CT	80.3	31.9	81.9	45.8	32.1	57.7	**69.4**	58.7	72.2	37.4	82.6	63.2	47.6	85.1	65.1	**89.5**	87.4	44.9	46.8	62.1
HRDA* + B^{3}CT	69.4	65.4	89.3	59.4	43.5	**57.9**	57.4	**67.4**	75.5	60.1	67.8	69.4	47.2	82.9	70.3	76.3	87.3	55.6	**63.3**	66.6
MIC* + B^{3}CT	**91.9**	**69.8**	**89.8**	**59.7**	**45.5**	56.8	52.1	65.8	**76.1**	**60.3**	86.5	**72.0**	**48.0**	**89.3**	73.3	81.5	**90.3**	**55.8**	62.1	**69.8**

Reply: Thank you for this insightful comment. We agree that quantifying the training overhead is a crucial aspect for a complete evaluation of our method. To address this, we have conducted a detailed analysis using the strong MIC model as our baseline on a single NVIDIA Titan RTX GPU.

We have added a new table (Table 8) to the revised manuscript to present these findings transparently. Here is a summary of the results:

Training Runtime: The training time per step increases from 5.24 s/step for the baseline MIC model to 8.51 s/step for our MIC + B^{3}CT method. This moderate increase is an expected trade-off, as our method incorporates an additional forward and backward pass for the hybrid branch to enable the cross-domain feature alignment that leads to significant performance gains.

GPU Memory Usage: The baseline MIC model consumes approximately 22,404 MB of VRAM. In contrast, our MIC + B^{3}CT implementation utilizes mixed-precision training that consumes only 19,334 MB. This paradoxical benefit arises because the architecture of B^{3}CT is amenable to stable mixed-precision training. This not only makes our method more memory-efficient but also enhances its

Table 4: **Training Overhead Comparison (MIC baseline on Titan RTX).**

Method	Training Time (s/step)	Peak GPU Memory (MB)	mIoU (GTAV→CS)
MIC (Baseline)	5.24	22,404	75.9%
MIC + B^3CT	8.51	19,334	76.4%

accessibility for training on standard 24GB GPUs.

In summary, B^3CT introduces a moderate increase in training duration in exchange for substantial improvements in segmentation accuracy. Crucially, this comes with the added advantage of a reduced memory footprint during training and, as noted by the reviewer, zero additional cost at inference. We believe this is a very favorable trade-off.

We have updated the manuscript to include this detailed analysis. Thank you again for prompting this valuable clarification.

References

[1] Y. Fu, R. Wang, Y. Fu, D. P. Paudel, and L. Van Gool, "Cross-view multi-modal segmentation@ ego-exo4d challenges 2025," *arXiv preprint arXiv:2506.05856*, 2025.

[2] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 289–305.

[3] W. Tranheden, V. Olsson, J. Pinto, and L. Svensson, "Dacs: Domain adaptation via cross-domain mixed sampling," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1379–1389.

[4] Y. Cheng, F. Wei, J. Bao, D. Chen, F. Wen, and W. Zhang, "Dual path learning for domain adaptation of semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9082–9091.

[5] N. Araslanov and S. Roth, "Self-supervised augmentation consistency for adapting semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 384–15 394.

[6] Q. Wang, D. Dai, L. Hoyer, L. Van Gool, and O. Fink, "Domain adaptive semantic segmentation with self-supervised depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8515–8525.

[7] P. Zhang, B. Zhang, T. Zhang, D. Chen, Y. Wang, and F. 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*, 2021, pp. 12 414–12 424.

[8] L. Hoyer, D. Dai, and L. 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.

[9] ——, "Hrda: Context-aware high-resolution domain-adaptive semantic segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 372–391.

[10] R. Gong, Q. Wang, M. Danelljan, D. Dai, and L. Van Gool, "Continuous pseudo-label rectified domain adaptive semantic segmentation with implicit neural representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7225–7235.

[11] L. Hoyer, D. Dai, H. Wang, and L. Van Gool, "Mic: Masked image consistency for context-enhanced domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 721–11 732.

[12] X. Huo, L. Xie, W. Zhou, H. Li, and Q. Tian, "Focus on your target: A dual teacher-student framework for domain-adaptive semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 027–19 038.

[13] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2517–2526.

[14] C. Sakaridis, D. Dai, and L. Van Gool, "Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3139–3153, 2020.

[15] X. Wu, Z. Wu, H. Guo, L. Ju, and S. Wang, "Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 769–15 778.

B^3CT: Three-branch Learning with Unlabeled Target Signals for Domain-Robust Semantic Segmentation

Chen Liang[1,2,3], Xin Zhao[1], Jian Jia[2,3], Junyan Wang[4], Lijun Cao[5], Jianguo Zhang[6], Weihua Chen[7]

[1]School of Computer and Communication Engineering, University of Science and Technology Beijing, No. 30 Xueyuan Road, Beijing, 100083, China.
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, No.19(A) Yuquan Road, Beijing, 100049, China.
[3]Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road, Beijing, 100190, China.
[4]Australian Institute for Machine Learning, The University of Adelaide, AIML Building, Lot Fourteen, Cnr North Terrace & Frome Road, Adelaide, 5000, Australia.
[5]Jiangsu Jiyuan Medical Technology Co., Ltd., No. 59 Meilan East Road, Taizhou, 610101, China.
[6]Research Institute of Trustworthy Autonomous Systems and Department of Computer Science and Engineering, Southern University of Science and Technology, No. 1088 Xueyuan Avenue, Shenzhen, 518055, China.
[7]DAMO Academy, Alibaba Group, Building 9, Area 4, Wangjing East Garden, Beijing, 100006, China.

Contributing authors: liangchen2022@ia.ac.cn; xinzhao@ustb.edu.cn; jianjia@outlook.com; junyan.wang@adelaide.edu.au; ljcao@foxmail.com; zhangjg@sustech.edu.cn; kugang.cwh@alibaba-inc.com;

Abstract

Semantic segmentation models often suffer from significant performance degradation when applied to unseen domains due to domain shifts. To address this challenge, we explore how to leverage unlabeled target-domain images during training to improve model robustness and generalization. Existing approaches primarily focus on achieving global alignment between source and target distributions, yet pay little attention to where and when such alignment should occur within the network. Through empirical observations, we find that different semantic contents are naturally aligned at different stages, and that alignment should be progressively enhanced as the quality of pseudo labels improves over training. Based on these insights, we propose a Three-Branch Coordinated Training (B^3CT) framework. In addition to conventional source and target branches, B^3CT introduces a dedicated alignment branch, where a hybrid-attention mechanism is used to guide feature-level consistency. To dynamically control the alignment strength, we design an Adaptive Alignment Controller (AAC)

and a coordinate weighting strategy that modulates the alignment intensity according to the training progress. Extensive experiments on GTAV→Cityscapes and SYNTHIA→Cityscapes benchmarks demonstrate that our method achieves competitive performance and exhibits strong robustness to domain shifts.

Keywords: semantic segmentation, feature alignment, cross-attention, domain generalization, adaptive training

1 Introduction

Deep neural networks have achieved remarkable success in various visual recognition tasks, but their generalization performance often degrades significantly when deployed in environments with domain shifts. This issue becomes particularly pronounced in dense prediction tasks such as semantic segmentation, where pixel-level accuracy is sensitive to changes in input distributions. As a result, developing segmentation models that are robust to domain shifts—without requiring additional annotations—has become an important research direction for both academia and industry. Such models are especially valuable in real-world applications where collecting pixel-wise labeled data for every new environment is prohibitively expensive.

To improve model generalization under distributional shifts, numerous studies have explored ways to learn domain-invariant representations. These approaches typically introduce alignment strategies at different levels of the model architecture to reduce discrepancies between source and target domains. Pixel-level alignment methods [1–6] employ image translation frameworks (e.g., GANs [7]) to harmonize visual styles across domains and alleviate low-level appearance discrepancies. Prototype-level methods [8–13] reduce semantic gap by aligning class-level features, promoting consistency in intermediate representations. Label-space alignment [14–18] encourages consistency in output predictions, often by minimizing entropy or enforcing structural regularities. Meanwhile, adversarial learning [14, 19, 20] has been widely adopted to implicitly align feature distributions via domain discriminators, and has shown strong empirical results when properly regularized.

Despite significant progress, two fundamental questions remain under-explored: **where** and **when** should alignment be performed? Regarding the former, some works [10, 11, 21] focus solely on aligning features at high-level layers, while others [22–24] treat all layers equally. We argue that this static design overlooks the hierarchical nature of visual information: low-level cues (e.g., textures, edges) and high-level semantics require alignment at different stages. A rigid, layer-invariant alignment strategy may lead to suboptimal feature fusion and hinder the model's ability to learn rich representations under complex domain shifts. Thus, adaptive control over where to align is essential for learning context-aware, semantically robust features.

In this paper, we propose a **Three-Branch Coordinated Training (B^3CT)** framework to enhance segmentation generalization under domain shifts. Beyond the conventional main and auxiliary training paths, B^3CT introduces a dedicated alignment branch, which leverages a hybrid-attention mechanism to enable inter-domain feature fusion through both self- and cross-attention. This alignment branch is designed to explicitly encode cross-domain relationships at different feature scales, serving as a dynamic bridge between source and target distributions. To dynamically control which features should be aligned at different network stages, we further introduce an **Adaptive Alignment Controller (AAC)** that selectively activates token-level alignment across layers, enabling stage-specific feature recalibration.

Additionally, the timing of alignment plays a crucial role in model stability and convergence. Over-reliance on early-stage features from auxiliary signals can introduce noise due to inaccurate pseudo-labels, while delayed alignment may reinforce source-domain bias and lead to underfitting on the target domain. To address this, we propose a **coordinate weighting strategy** based on pseudo accuracy estimates, which gradually adjusts the contribution of the alignment branch during training. This curriculum-inspired

approach encourages the model to initially rely on reliable source supervision and progressively incorporate target signals as pseudo-label quality improves.

Our contributions are summarized as follows:

- We propose a three-branch coordinated training framework (B^3CT) that jointly addresses **where** and **when** to align in a unified setting, aiming to improve generalization under domain shifts while maintaining training stability.
- We design an Adaptive Alignment Controller (AAC) to dynamically select alignment locations across different network stages, enabling feature-level adaptivity tailored to hierarchical representations.
- We introduce a coordinate weighting mechanism to schedule alignment progression based on pseudo-label reliability, bridging the gap between early robustness and late-stage accuracy.
- We demonstrate that our method achieves competitive performance (76.4% on GTA5 $\rightarrow$ Cityscapes and 68.7% on SYNTHIA $\rightarrow$ Cityscapes), showcasing its robustness to domain shifts and compatibility with transformer-based segmentation backbones.

2 Related Work

2.1 Semantic Segmentation

Semantic segmentation aims to assign category labels to every pixel in an image, enabling a fine-grained understanding of visual scenes. Since the advent of deep learning, numerous methods have been proposed based on the fully convolutional network (FCN) architecture [25], which serves as the foundational framework for modern semantic segmentation. To mitigate the limitations of limited receptive field and information loss across network stages, researchers have enhanced FCN-based models in various ways, such as by refining contextual information [26–28], enlarging receptive fields through spatial pyramids [29, 30], or incorporating attention mechanisms [31–33]. More recently, Vision Transformers (ViTs) [34] have demonstrated strong potential in semantic segmentation by leveraging their long-range dependency modeling capabilities. Models like Swin Transformer [35] further introduce hierarchical attention structures for better spatial reasoning. Despite these advancements, segmentation models still face challenges in generalizing to unseen domains due to substantial distribution shifts.

2.2 Semantic Segmentation under Domain Shifts

To improve model robustness across diverse environments, many recent works address the problem of semantic segmentation under domain shifts by encouraging the learning of domain-invariant representations. One common line of research focuses on pixel-level alignment [3–6, 36], where image translation techniques such as CycleGAN [7] are used to unify visual styles across different domains prior to segmentation. Another strategy is prototype-level alignment [8–13], which aligns class-specific feature centers between domains. Output-space alignment approaches [14–18] match prediction distributions via probability consistency and entropy minimization, while consistency-based strategies also explore feature alignment across multiple network views [37]. Additionally, adversarial learning frameworks [14, 19, 20] utilize domain discriminators to guide the alignment of feature distributions. Self-training techniques [38–42] leverage pseudo-labels generated on unlabeled samples to iteratively improve model performance. These approaches share the common goal of enabling models to generalize better across domains without requiring annotations in unseen environments.

2.3 Attention Mechanisms in Cross-Domain Learning

The self-attention mechanism, first introduced in the Transformer architecture [43], has become a powerful tool for modeling contextual dependencies in vision tasks. ViT [34] pioneered the use of self-attention in visual domains by representing images as token sequences, enabling long-range interaction. Swin Transformer [35] further optimized computational efficiency by introducing a hierarchical window-based attention structure. In cross-domain settings, attention mechanisms have also been explored to promote feature fusion and alignment. Cross-attention mechanisms [24, 44, 45] allow features from different domains or views to interact explicitly. For instance, Kang

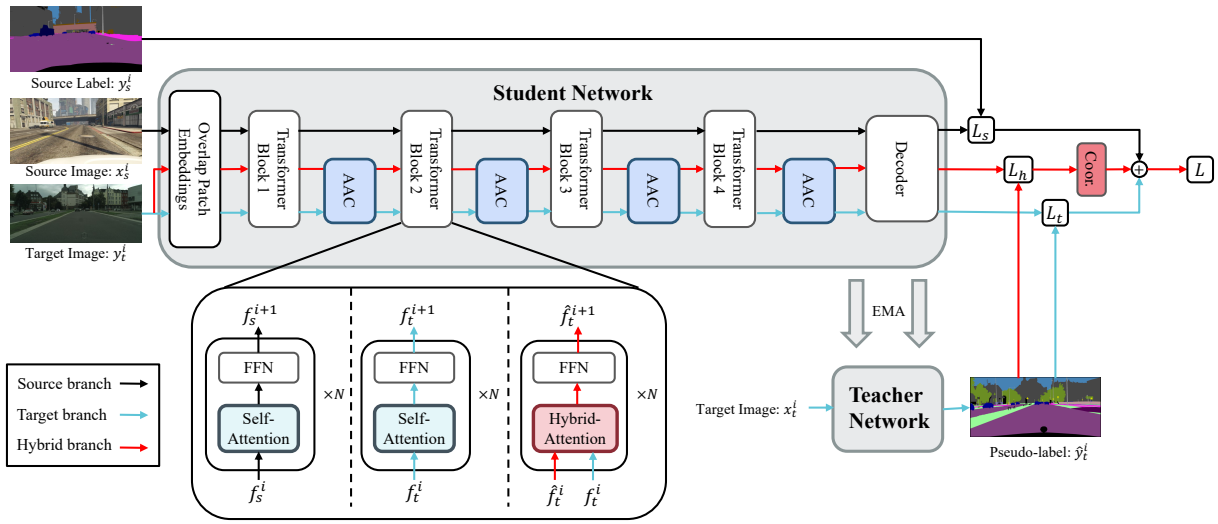


Fig. 1: **Illustration of the B^{3}CT framework.** The B^{3}CT framework is mainly divided into two parts: a student network and a teacher network. The teacher network is updated via exponential moving average (EMA) of the student network weights and provides stable training signals for images from auxiliary domains. The student network consists of three parallel branches with shared weights, each utilizing a distinct attention mechanism. A coordinate weighting strategy is employed to dynamically adjust the contribution of each branch to the final loss.

[46] employed self-attention on region features to capture semantic and spatial relationships, while Chen [44] introduced cross-attention between multi-scale patches to improve feature integration. Xu [24] applied cross-attention to mitigate noisy supervision in classification tasks under domain shifts. Inspired by these ideas, our work incorporates a hybrid-attention design to facilitate effective alignment across domains within a coordinated training framework, aiming to enhance generalization under domain shift.

3 The Proposed Method

In this section, we present the details of our B^{3}CT framework, as shown in Fig. 1. The framework consists of three parallel training branches that share weights and are optimized end-to-end. In the third branch, a modified attention module named *hybrid-attention* is used to promote robust feature consistency across domains. This module integrates both intra-domain self-attention and inter-domain cross-attention to enhance contextual understanding and improve cross-domain representation learning.

To further refine the alignment behavior across network layers, we introduce an *Adaptive Alignment Controller* (AAC), which dynamically adjusts alignment intensity based on the semantic hierarchy of attention blocks. Finally, we propose a *coordinate weighting mechanism* to regulate the relative influence of the hybrid-attention branch during training, based on the evolving quality of predictions. Together, these components form a unified architecture that improves segmentation performance under domain shifts by enhancing the model's generalization capacity.

3.1 The Three-Branch Coordinated Training Framework

Given a labeled source domain dataset $\mathbb{D}_s = \{(x_s^{(i)}, y_s^{(i)})|y_s^{(i)} \in \mathbb{R}^{H\times W}\}_{i=1}^{N_s}$ and an auxiliary unlabeled dataset $\mathbb{D}_t = \{x_t^{(i)}\}_{i=1}^{N_t}$ representing samples from a different domain, the goal is to train a segmentation model that achieves robust pixel-level predictions under domain shifts. Here, N_s and N_t denote the number of samples in the source and auxiliary sets respectively, while H and W indicate the height and width of each image.

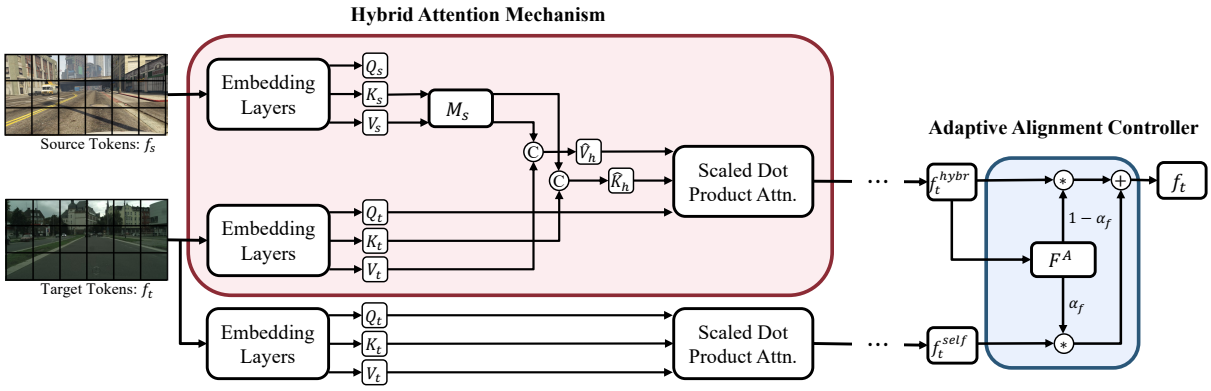


Fig. 2: **Illustration of the hybrid-attention mechanism and AAC.** Taking the feature tokens f_s^i, f_t^i from source and target domains as inputs, three embedding layers project these tokens to the corresponding query Q_i, key K_i, and value tokens V_i respectively, where $i \in \{s, t\}$. In hybrid-attention, query tokens Q_t are from target domain, key K_i and value tokens V_i are formed by $[K_s; K_t]$ and $[V_s; V_t]$. At each stage of the Transformer[47], the Adaptive Alignment Controller (AAC) dynamically adjusts the weights of the sum of features derived from two attention mechanisms through the learning of a relative attention mask α_f.

To achieve this, we follow a teacher-student training paradigm inspired by prior works [8, 38, 41, 42, 48], where the teacher network provides stable training signals without requiring annotations in $\mathbb{D}_t$. This approach supports an end-to-end learning process while avoiding iterative optimization procedures.

Our framework consists of three parallel branches that share weights: a *source branch*, a *target branch*, and a *hybrid branch*. In the source branch, labeled source images are fed into a self-attention-based transformer to obtain per-pixel semantic predictions. The standard cross-entropy loss is applied to supervise each image $x_s^{(i)}$ from $\mathbb{D}_s$:

$$L_s^i = -\frac{1}{HW}\sum_{j=1}^{HW}\sum_{c=1}^{C} y_s^{i,j,c} \log g(x_s^{(i)})^{(j,c)} \qquad (1)$$

where C is the number of semantic categories. g represents the forward function of the student network using self-attention only.

We follow a consistency-based training pipeline where auxiliary-domain samples $\mathbb{D}_t$ are utilized without additional labels. The teacher network provides online predictions $\hat{y}_t$ as training signals for unlabeled data. After filtering predictions using a confidence threshold, the cross-entropy loss for the auxiliary data is formulated as:

$$L_t^i = -\frac{1}{HW}\sum_{j=1}^{HW}\sum_{c=1}^{C} \hat{y}_t^{i,j,c} \log g(x_t^{(i)})^{(j,c)} \qquad (2)$$

To support effective cross-domain representation learning, we build upon the mean teacher model [48, 49] and introduce an additional branch that jointly processes both source and auxiliary-domain images in a hybrid-attention mechanism. As shown in Fig. 1, the resulting *hybrid branch* shares weights with the other branches but leverages distinct attention operations for enhanced representation robustness.

Given a pair of augmented source and auxiliary-domain images, the student network first downsamples and reshapes the features into a sequence of tokens $f_s \in \mathbb{R}^{N\times d}$ and $f_t \in \mathbb{R}^{N\times d}$, where each token is a vector representation of a local spatial patch after patch embedding, following the standard formulation in Vision Transformers. Each token corresponds to one spatial location and serves as the basic unit for attention-based interactions. Here, $N = \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}$ is the number of tokens in the i-th stage, and d is the number

of feature channels. These tokens are then passed through the hybrid-attention layers to propagate and integrate information across domains.

Inspired by recent efforts in multi-source attention and multi-modal fusion [44, 50], we design a hybrid-attention mechanism that integrates both intra-domain self-attention and inter-domain cross-attention to enable more flexible cross-domain interactions. Instead of enforcing explicit category-level correspondence between source and target tokens, the hybrid-attention module exposes representations from both domains to a shared attention space, allowing the model to automatically learn beneficial correlation patterns during training. The key and value features from the two domains are concatenated as $K_h = [K_s; K_t]$ and $V_h = [V_s; V_t]$, and the target-domain query Q_t computes attention over these combined representations:

$$Attn_{hybrid}(Q_t, K_s, K_t, V_s, V_t) = Softmax(\frac{Q_t[K_s; K_t]^\top}{\sqrt{d_k}})[V_s; V_t] \quad (3)$$

Since there are both self-attention ($Q_s K_s^\top$ and $Q_t K_t^\top$) for intra-domain feature fusion and cross-attention ($Q_s K_t^\top$ or $Q_t K_s^\top$) for inter-domain feature fusion, we call it the hybrid-attention module.

Furthermore, before aligning at the feature level, because the source and target domain data have different distributions in the feature space, their queries, keys, and values also vary in basis for calculating attention. When the key of the source and target domains are concatenated together, the network is forced to map the source and target domains on the same basis. In other words, for any query extracted by the network in any domain, it is completely equivalent to calculating the similarity under the basis of the source domain and the target domain. When the network learns this equivalence relationship, it achieves alignment between the two domains at the feature level. However, it is difficult for the network to directly learn unified feature representation ideally [51]. Therefore, in the hybrid branch, a linear layer M_s is used to perform quadratic mapping on the key and value of the source domain to the target domain.

$$Attn_{hybrid}(Q_t, K_s, K_t, V_s, V_t) = Softmax(\frac{Q_t[M_s^K(K_s); K_t]^\top}{\sqrt{d_k}})[M_s^V(V_s); V_t] \quad (4)$$

This way, even if the features of different domains cannot be fully aligned, the mapped query and key can be guaranteed to be on the same basis. It is noteworthy that the linear layer M_s is exclusively employed in the hybrid branch. The parameters and computations of both the source branch and target branch remain unaltered. This ensures that the inference process remains consistent with any baseline, without introducing additional computational complexity. Then we have the cross-entropy loss on the hybrid branch:

$$L_h^i = -\frac{1}{HW}\sum_{j=1}^{HW}\sum_{c=1}^{C}\hat{y}_t^{i,j,c}\log g^{(h)}(x_t^{(i)})^{(j,c)} \quad (5)$$

where $g^{(h)}$ represents the forward function of the student Transformer network using hybrid-attention.

To further explore the feature alignment capability of the hybrid branch, in the following section, we introduce the proposed AAC, where dynamic alignment adjustment can be made based on specific semantic content.

3.2 Adaptive Alignment Controller

While hybrid-attention facilitates feature-level interaction across domains, applying such integration uniformly across all stages may not be optimal [52]. In deep architectures, earlier layers tend to focus on low-level cues such as texture or edges, whereas deeper layers capture high-level semantic structures. We thus posit that selective and stage-aware integration is essential: for example, visual concepts like "road" may benefit from earlier integration due to shared visual properties across domains, while abstract objects such as "vehicles" require higher-layer integration where semantic context dominates.

To address this, we propose the Adaptive Alignment Controller (AAC), which enables dynamic control over feature integration at different semantic levels of the network. Inspired by

observations from prior work [53], which emphasizes the varying domain informativeness of features at different depths, AAC refines the integration granularity by learning a spatially-aware fusion mask. Unlike fixed-weight schemes, our method allows fine-grained control of representation fusion across spatial and semantic axes.

As shown in Fig. 2, during training in the hybrid branch, features are independently computed using self-attention (f_t^{self}) and hybrid-attention (f_t^{hybr}). A learned dense mask is then applied to blend these representations at each network stage. The attention decoder F^A generates a pixel-level fusion map $a_f = \sigma(F^A(f_t^{hybr})) \in [0,1]^{H\times W\times C}$, where the sigmoid activation constrains values between 0 and 1. The final representation is obtained by a weighted sum:

$$f_t = a_f \odot f_t^{hybr} + (1 - a_f) \odot f_t^{self} \tag{6}$$

This design equips our model with the ability to adaptively emphasize different information pathways depending on the stage and spatial location, effectively tailoring the representation to cross-domain variability. Compared to prior approaches that enforce domain invariance through adversarial training, our approach offers more nuanced and context-aware representation adjustment. Particularly in dense prediction tasks like semantic segmentation, such granularity is essential to preserve spatial detail and semantic integrity. AAC thus serves as a key enabler for robust generalization across diverse domains by modulating attention-based fusion with precision.

Furthermore, AAC serves as an additional safeguard: when the hybrid-attention module produces noisy or semantically incorrect correlations between source and target tokens, AAC automatically downweights the hybrid features and relies more on self-attention. This prevents harmful cross-category alignment and stabilizes the overall training dynamics.

3.3 Coordinate weight

Previous approaches often rely on an iterative framework in which a model pre-trained on source data generates pseudo-labels for auxiliary data, followed by separate training stages. In contrast, our framework adopts an end-to-end training paradigm based on the mean teacher model [54], where the student network is updated via standard backpropagation and the teacher network is updated as the exponential moving average (EMA) of the student weights at each step t. This strategy promotes stability and consistency in model predictions during training.

Building on this foundation, our B^{3}CT framework jointly processes both labeled and unlabeled data streams across three parallel branches. A key challenge, however, lies in determining how to regulate the contribution of the hybrid-attention branch throughout training. Early integration of hybrid attention may introduce noise due to initially unstable predictions on unlabeled inputs, thereby hindering the learning of discriminative representations. Conversely, delaying this integration may cause the model to overfit to source-specific patterns, reducing its generalization ability.

Instead of relying on pre-defined stage scheduling or auxiliary classifiers [55, 56], we propose a confidence-aware coordination mechanism based on the predictive consistency between the student and teacher networks. Specifically, we use the pseudo-accuracy $Acc(\bar{y}_t, \hat{y}_t)$—computed between the student's prediction p_t and the teacher's output $\hat{y}_t$—to dynamically control the hybrid branch's influence. This metric reflects the model's confidence in auxiliary-domain representations and evolves throughout training. Formally, the coordinate weight is defined as:

$$Coor(p_t, \hat{y}_t) = Acc(\bar{y}_t, \hat{y}_t) \times (1 - e^{-iter\cdot\alpha}),$$
$$\text{where } \bar{y}_t^{i,j} = \underset{k}{\operatorname{argmax}}\, p_t^{i,j,k} \tag{7}$$

Here, *iter* denotes the current training iteration and α controls the update pace. The coordinate weight thus modulates the overall loss dynamically:

$$L = L_s + L_t + Coor(p_t, \hat{y}_t) \cdot L_h \tag{8}$$

In our experiments, we employ the HRDA architecture as both student and teacher models. B^{3}CT serves as the foundational framework for further analysis.

Table 1: **Component Ablation for B^{3}CT** (GTAV→Cityscapes, HRDA backbone).

hybrid branch	coor.weight	AAC	mIoU
–	–	–	73.8 ± 0.1
✓	–	–	74.2 ± 0.2
✓	✓	–	74.3 ± 0.2
✓	–	✓	74.4 ± 0.2
✓	✓	✓	**74.8** ± 0.1

Table 2: **Different data flow in hybrid branch** (GTAV→Cityscapes, HRDA backbone).

source	target	mIoU
–	–	73.8
✓	–	73.1
–	✓	**74.8**
✓	✓	74.5

Table 3: **Ablation study of the hybrid-attention with AAC in each stage. (GTAV→Cityscapes, HRDA backbone)** The first line shows the performance without the third branch.

Stage1	Stage2	Stage3	Stage4	mIoU
–	–	–	–	73.8
✓	–	–	–	74.3
–	✓	–	–	74.3
–	–	✓	–	74.4
–	–	–	✓	74.4
✓	✓	–	–	74.3
–	✓	✓	–	74.5
–	–	✓	✓	74.6
✓	✓	✓	–	74.5
–	✓	✓	✓	74.6
✓	✓	✓	✓	**74.8**

Table 4: **Ablation on the linear mapping. (GTAV→Cityscapes, HRDA backbone)**

M_s	M_t	mIoU
–	–	74.5
✓	–	74.3
–	✓	**74.8**
✓	✓	**74.8**

It is worth noting that this coordination mechanism exclusively modulates the contribution of the hybrid branch, while losses on the labeled and auxiliary data branches are computed with fixed weighting. In the HRDA training protocol, pseudo-label quality is already regulated by a confidence threshold T applied to softmax scores. However, since the hybrid branch performs uniform attention over all tokens, a time-sensitive, confidence-guided weight provides a more nuanced solution to the "when to integrate" problem. Rather than applying a binary threshold to activate or deactivate hybrid interaction, our design employs a smooth transition via a gradually increasing weight—leading to improved stability and generalization throughout training.

4 Experiments

Datasets: We use the GTAV dataset [57] and the Synthia dataset [58] for the source domain. The GTAV dataset consists of 24,966 synthetic images with resolution 1914×1052. The Synthia dataset contains 9,400 synthetic images with resolution 1280×760. For the target domain, we use the Cityscapes street scene dataset [59], which contains 2,975 training and 500 validation images with resolution 2048×2048.

Implementation Details: We use HRDA [49] as our default network, which consists of an MIT-B5 encoder [47], a context-aware feature fusion decoder [48] and a lightweight SegFormer MLP decoder [47] as the scale attention decoder. The self-training strategy and training parameters are all the same as HRDA's settings. We use AdamW with a learning rate of 6×10^{-5} for the encoder and 6×10^{-4} for the decoder, linear learning rate warmup, and DACS [60] data augmentation. The batch size is set as 2. The experiments are conducted on a Titan RTX GPU with 24GB memory. The ablation experiments were

Table 5: Ablation study for AAC. (GTAV→Cityscapes, HRDA backbone) The AAC is compared with no feature fusion strategy and simple average pooling.

Feature fusion	mIoU
–	74.3
Average Pooling	74.4
Fused Attention	**74.8**

conducted on the GTAV → Cityscapes benchmark, while all of the results are averaged over 3 random seeds.

4.1 Ablation studies for B^{3}CT

Component ablation for B^{3}CT. The B^{3}CT self-training framework is composed of three key components: the hybrid-attention branch, the Adaptive Alignment Controller (AAC) for selective representation fusion, and the coordinate weight mechanism for dynamically adjusting the branch's training contribution. We conduct a detailed ablation study on these components, as summarized in Tab. 1.

The coordinate weight regulates the overall integration pace of the hybrid branch during training, while AAC adjusts the fusion behavior at a finer granularity—across semantic categories and network stages. Simply applying coordinate weighting across all stages can result in suboptimal alignment or even training conflicts, as observed in Row 3 of Tab. 1. However, by jointly using AAC and coordinate weight (Row 5), the network can dynamically learn both where and when to perform domain-robust representation fusion, leading to a clear performance gain. In particular, while introducing the hybrid branch alone yields a 0.4% mIoU improvement, integrating both AAC and the coordination mechanism brings an additional 0.6% mIoU increase.

Importantly, AAC and coordinate weight exhibit a strong synergy. Without adaptive weighting, the AAC struggles to converge to effective fusion strategies due to early-stage uncertainty. Conversely, using only coordinate weighting without semantic-aware control fails to differentiate fusion behavior across layers and categories. Together, they provide complementary control over the training dynamics and significantly boost model generalization.

Different data flows in the hybrid branch. The B^{3}CT framework includes three parallel branches: a source-only branch, a target-only branch, and a hybrid branch that integrates information across domains. In the hybrid branch, we experiment with three types of input data flow: (1) source-only, (2) source + target, and (3) target-only. For the first variant, hybrid-attention and AAC are computed using queries from source tokens, while key and value tokens are drawn from both domains. Only the source prediction is used for backpropagation. In the second variant, both domains provide queries and yield dual predictions. In the third, only target queries are used, and only target predictions are optimized.

Results in Tab. 2 reveal that using only target data in the hybrid branch leads to the best performance. This demonstrates that enhancing representations directly on the target domain leads to stronger generalization, aligning with our goal of achieving high performance under domain shift. Interestingly, the source+target flow also outperforms the source-only variant, confirming the effectiveness of the proposed hybrid-attention and AAC mechanisms in leveraging cross-domain cues for robust feature learning.

Hybrid-attentionin different stages. In order to verify that the alignment of each stage is necessary, in Tab. 3, we demonstrate the effect of hybrid-attention in different Transformer stages. Since the alignment processes in the four stages increase, the adaptation performance of our method arises from 73.8% to 74.8%. The experimental results indicate that aligning features in each layer of the network makes sense. More hybrid-attention are applied, the tighter the fusion of features between the source and target domains, the more it helps the model learn domain-invariant features. It is worth noting that in this experiment, all the hybrid-attentions are together with AAC.

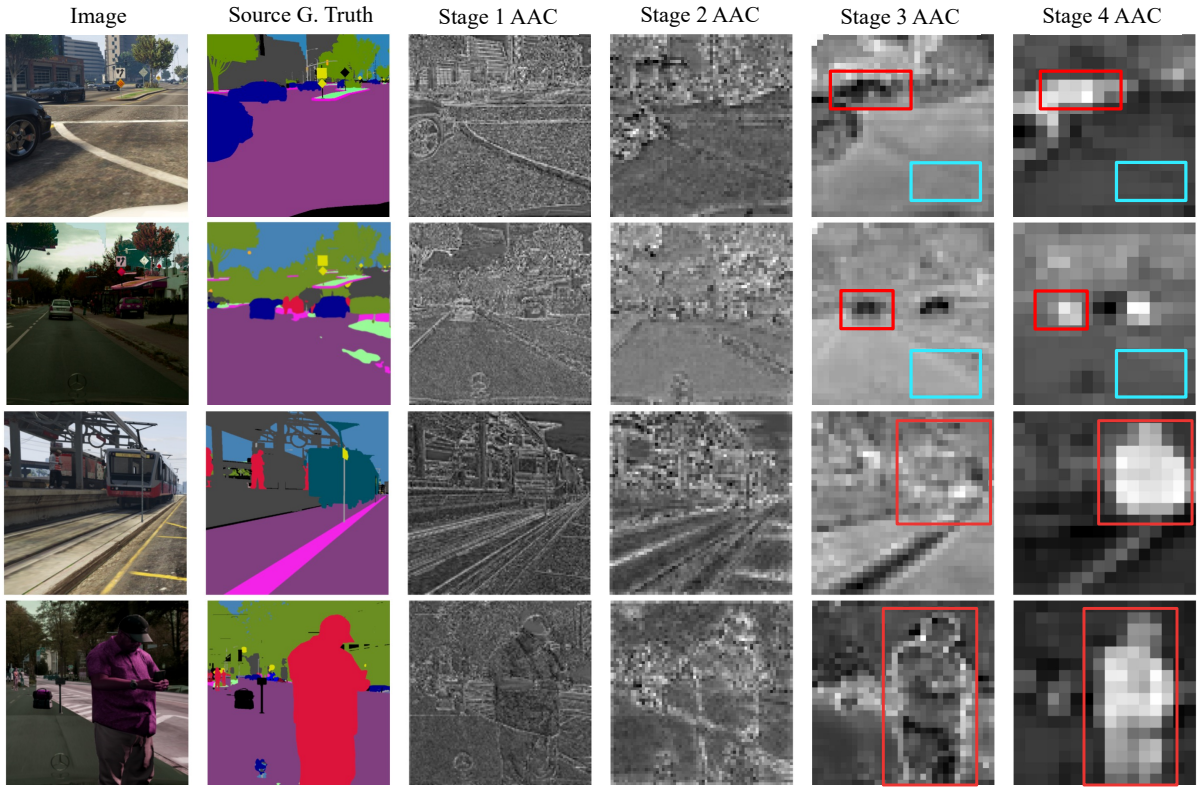


Fig. 3: **Visual examples of the AAC in different Transformer[47] stage.** The brighter the color, the greater the feature weight calculated using hybrid-attention, while the darker the color, the greater the feature weight calculated using self-attention. The AAC learns to decide whether features from two domains should be aligned at certain stages.

Ablation on the linear mapping M_s. We further investigate different configurations of where the linear mapping is applied in the hybrid-attention module. As shown in Tab. 4, applying the mapping only to the source-domain features achieves the highest accuracy. Mapping both domains or mapping the target-domain features leads to inferior results, confirming the effectiveness of the asymmetric design.

4.2 Ablation studies for the AAC

Quantitative experiments on AAC. More specifically on the AAC module, we conduct quantitative experimental verification on this module in Tab. 5. If only the hybrid-attention is conducted without adding AAC, the final model mIoU is 74.3%. Given that the function implemented by AAC is the dynamic weighting of two feature maps, we attempt to directly average the two feature maps and result in no improvement. However, Applying AAC will result in an improvement of 0.42% mIoU. It can be seen that directly using hybrid-attention for feature alignment can easily lead to suboptimal results due to domain gaps. Using AAC to adaptively align and adjust each token at each stage can achieve the optimal alignment effect.

Qualitative experiments on AAC. Furthermore, in Fig. 3, we visualized the alignment strategies learned by the AAC module at each stage of the network. In the figure, brighter colored tokens represent AAC giving greater weight to the features calculated by hybrid-attention; On the contrary, darker tokens represent AAC focusing more on self-attention features. In other words, a brighter color indicates that the network is more

Table 6: **Comparison with exsiting methods on the GTAV/Synthia to Cityscapes benchmarks.** The **first** and second highest scores are represented by bold font and underline respectively. * represents results from our re-implementation. † represents methods use additional clear-weather geographically-aligned reference images.

	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU
GTAV → Cityscapes																				
CBST [38]	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 [60]	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 [5]	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 [42]	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 [61]	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 [8]	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 [48]	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
HRDA [49]	96.4	74.4	91.0	61.6	51.5	57.1	63.9	69.3	91.3	48.4	94.2	79.0	52.9	93.9	84.1	85.7	75.9	63.9	67.5	73.8
IR^{2}F-RMM [16]	97.5	80.0	91.0	60.0	53.3	56.2	63.9	72.4	91.7	51.0	94.2	79.0	51.1	94.3	84.7	86.7	75.9	62.6	67.8	74.4
MIC [62]	**97.4**	80.1	91.7	61.2	56.9	59.7	66.0	71.3	91.7	**51.4**	94.3	79.8	56.1	94.6	85.4	90.3	80.4	64.5	68.5	75.9
DTS [63]	97.0	80.4	91.8	60.6	**58.7**	61.7	**67.9**	73.2	92.0	45.4	94.3	**81.3**	**58.7**	**95.0**	**87.9**	**90.7**	**82.2**	65.7	69.0	**76.5**
DAFormer + B^{3}CT	96.5	75.2	90.7	60.6	45.5	57.5	63.7	68.6	90.9	48.4	91.9	76.7	49.2	92.1	59.5	68.3	64.6	58.7	67.9	69.8
HRDA + B^{3}CT	96.5	75.0	91.2	59.6	55.4	58.5	66.3	69.1	91.9	49.3	94.4	79.2	55.3	94.2	83.2	88.7	80.3	63.7	68.4	74.8
MIC + B^{3}CT	**97.4**	**81.1**	**92.1**	**63.5**	54.3	**62.0**	66.8	**74.2**	**92.1**	50.8	**94.6**	80.1	55.9	94.9	86.9	89.6	79.9	**67.0**	**69.3**	76.4
Synthia → Cityscapes																				
CBST [38]	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 [60]	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 [5]	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 [42]	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 [61]	**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 [8]	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 [48]	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
HRDA [49]	85.2	47.7	88.8	49.5	4.8	57.2	65.7	60.9	85.3	–	92.9	79.4	52.8	89.0	–	64.7	–	63.9	64.9	65.8
MIC [62]	86.6	50.5	89.3	47.9	7.8	59.4	**66.7**	**63.4**	87.1	–	94.6	81.0	58.9	**90.1**	–	61.9	–	67.1	64.3	67.3
IR^{2}F-RMM [16]	90.4	54.9	89.4	48.0	7.4	59.0	65.5	63.2	87.8	–	94.1	80.5	55.8	90.0	–	65.9	–	64.5	66.8	67.7
DTS [63]	89.1	54.9	89.0	39.1	8.7	**61.6**	67.4	64.3	88.8	–	94.0	**82.2**	**60.7**	89.6	–	62.6	–	**68.5**	64.9	67.8
DAFormer + B^{3}CT	82.8	44.2	86.7	38.9	5.0	55.0	63.0	61.2	83.6	–	82.9	76.3	47.9	88.0	–	58.6	–	61.8	65.3	62.6
HRDA + B^{3}CT	90.5	57.5	88.7	46.3	7.5	57.7	66.6	63.3	89.1	–	94.2	80.5	55.8	89.0	–	54.4	–	64.6	65.7	67.0
MIC + B^{3}CT	90.4	54.6	**90.6**	**57.3**	**11.4**	58.3	64.9	62.1	**89.5**	–	**94.8**	79.8	54.2	90.0	–	**67.9**	–	66.0	**67.4**	**68.7**
Cityscapes → ACDC																				
ADVENT [15]	72.9	14.3	40.5	16.6	21.2	9.3	17.4	21.2	63.8	23.8	18.3	32.6	19.5	69.5	36.2	34.5	46.2	26.9	36.1	32.7
MGCDA† [64]	73.4	28.7	69.9	19.3	26.3	36.8	53.0	53.3	75.4	32.0	84.6	51.0	26.1	77.6	43.2	45.9	53.9	32.7	41.5	48.7
DANNet† [65]	84.3	54.2	77.6	38.0	30.0	18.9	41.6	35.2	71.3	39.4	**86.6**	48.7	29.2	76.2	41.6	43.0	58.6	32.6	43.9	50.0
DAFormer [48]	58.4	51.3	84.0	42.7	35.1	50.7	30.0	57.0	74.8	52.8	51.3	58.3	32.6	82.7	58.3	54.9	82.4	44.1	50.7	55.4
HRDA* [49]	69.1	64.8	86.4	53.9	40.3	56.5	41.1	63.9	72.8	58.8	62.0	69.9	44.2	87.0	**73.5**	82.0	87.8	45.4	63.0	64.3
MIC* [62]	89.8	62.0	86.6	53.5	40.7	48.9	58.8	60.1	73.2	57.1	85.7	67.9	44.6	87.9	72.5	84.3	88.4	51.6	56.8	66.9
DAFormer* + B^{3}CT	80.3	31.9	81.9	45.8	32.1	57.7	**69.4**	58.7	72.2	37.4	82.6	63.2	47.6	85.1	65.1	**89.5**	87.4	44.9	46.8	62.1
HRDA* + B^{3}CT	69.4	65.4	89.3	59.4	43.5	**57.9**	57.4	**67.4**	75.5	60.1	67.8	69.4	47.2	82.9	70.3	76.3	87.3	55.6	**63.3**	66.6
MIC* + B^{3}CT	**91.9**	**69.8**	**89.8**	**59.7**	**45.5**	56.8	52.1	65.8	**76.1**	**60.3**	86.5	**72.0**	**48.0**	**89.3**	73.3	81.5	**90.3**	**55.8**	62.1	**69.8**

inclined to align the feature representations of the token across different domains certain stages.

Firstly, from stage 1, it can be seen that the network mainly aligns at texture and edges. This pattern is similar to using CycleGAN [7] to align images from different domains at the input level, such as making categories from different domains have the same texture and color information. As the network layer deepens, AAC increasingly tends to align at the level of semantic object features. For example, in Fig. 3, AAC aligns the road and other categories in stage 3, emphasized by the blue boxes, and aligns the vehicle and other categories in stage 4, emphasized by the red boxes. Besides, the features that have already been aligned in stage 3 no longer participate in the alignment in stage 4. AAC will automatically select which layer is the most suitable for alignment for each patch. These results are consistent with our motivation for designing AAC.

4.3 Ablation studies for coordinate weight

To explore the most suitable feature alignment time, we investigate the effects of the coordinate weight while the hybrid branch and AAC are retained. The hyperparameter α of coordinate weight in Eq. 7 is designed to achieve harmonious training among the three branches. 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. Among them, the orange horizontal line represents no adjustment for the participation level of the hybrid branch by

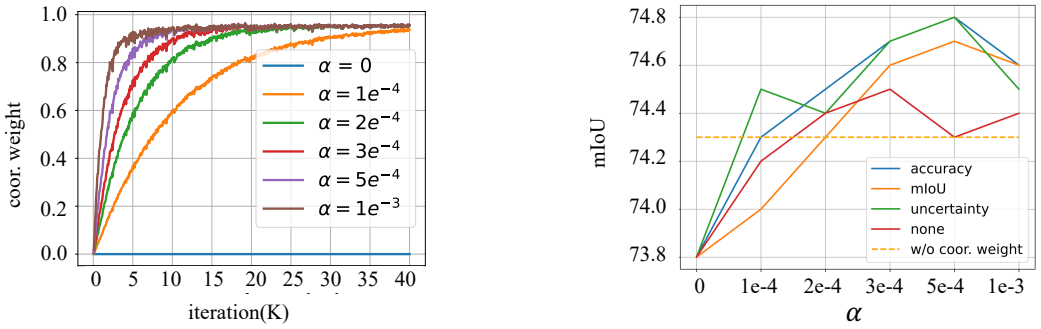


Fig. 4: Coordinate weight curves and performance comparison between various α. Setting $\alpha = 0$ indicates that the hybrid branch is not involved in the training. Our method reaches 74.4 without coordinate weight as shown in the dashed line.

removing coordinate weight, which means assigning it the same loss weight as other branches. Setting $\alpha = 0$ means that the hybrid branch is not involved and only the source branch and target branch are available for model training. As the α increases, the earlier the target domain is involved in the training, while the overall model effect shows a trend of first increasing and then decreasing.

In Eq. 7, $Acc(\bar{y}_t, \hat{y}_t)$ represents the performance of the student model. To find out the most appropriate evaluation, we try to replace the accuracy by mIoU [5], uncertainty [66], and a constant. The experimental results showed that both accuracy and uncertainty achieved the best results. For the convenience of calculation, we use accuracy as the product term in other experiments. We take $\alpha = 5e^{-4}$ as our default setting according to the experimental results.

4.4 Comparison with existing methods

To demonstrate the effectiveness and generalizability of the proposed method, we report the quantitative results in terms of mean Intersection-over-Union (mIoU, %) on two standard benchmarks in Tab. 6. For fair comparison on SYNTHIA→Cityscapes, we follow the convention of converting 13-category mIoU to the 16-category mIoU metric [63]. To further evaluate compatibility with existing architectures, we additionally integrate our B^{3}CT framework with the state-of-the-art model MIC [62].

On the GTAV→Cityscapes task, B^{3}CT achieves absolute gains of 1.5%, 1.0%, and 0.5% mIoU when built upon DAFormer, HRDA, and MIC respectively. Similarly, on SYNTHIA→Cityscapes, the improvements are 1.7%, 1.2%, and 1.4% mIoU. These consistent gains highlight the flexibility of B^{3}CT in enhancing various Transformer-based segmentation frameworks. To further assess the robustness of our approach under real-world distribution shifts, we evaluate on the clear-to-adverse-weather Cityscapes→ACDC benchmark. This setting involves transferring from clear daytime scenes (Cityscapes) to diverse adverse conditions including nighttime, fog, rain, and snow (ACDC). As shown in Tab. 6, B^{3}CT improves DAFormer from 55.4% to 62.1%, HRDA from 64.3% to 66.6%, and MIC from 66.9% to 69.8%. These results demonstrate that B^{3}CT not only enhances synthetic-to-real adaptation but also delivers strong generalization under severe and heterogeneous real-world domain shifts.

4.5 Qualitative Analysis

To visually illustrate the improvements brought by B^{3}CT, qualitative comparisons are shown in Fig. 5, using examples corresponding to Tab. 6. Compared to HRDA [49] and MIC [62], the incorporation of B^{3}CT into MIC (i.e., MIC + B^{3}CT) significantly improves the semantic coherence of the predicted segmentation masks.

As observed in Fig. 5, baseline models often struggle with fine-grained semantic categories, particularly under challenging lighting or

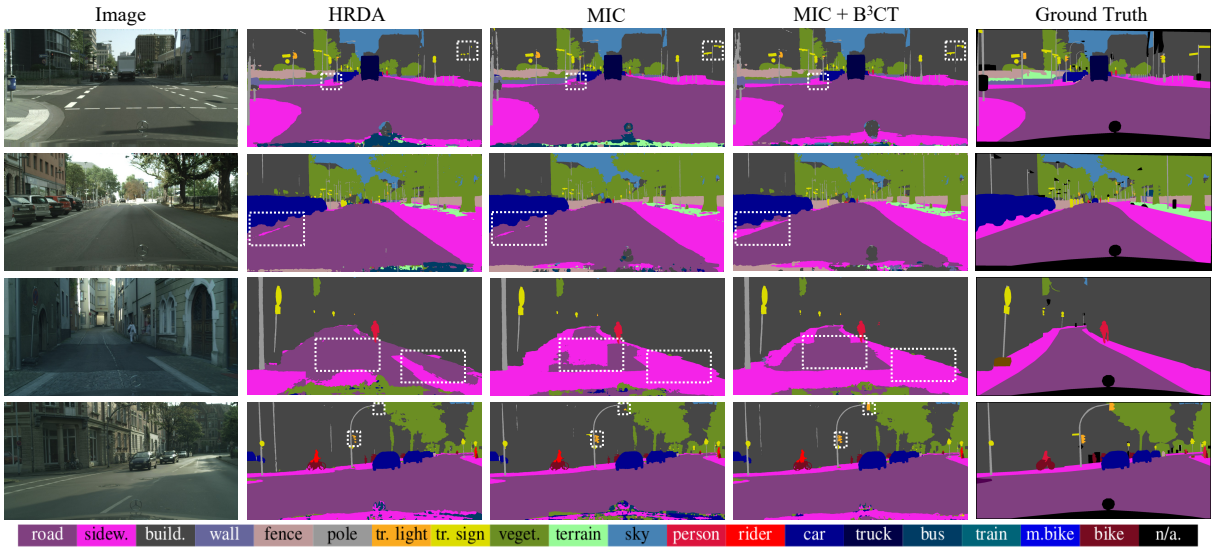


Fig. 5: Visualization of the segmentation results of previous work and B^{3}CT framework. The qualitative comparison are on the GTAV→Cityscapes(row 1 and 2) and Synthia→Cityscapes(row 3 and 4). The white dashed boxes indicate that the B^{3}CT framework has better semantic consistency, which can reduce the situation where a single object is segmented into different semantic categories.

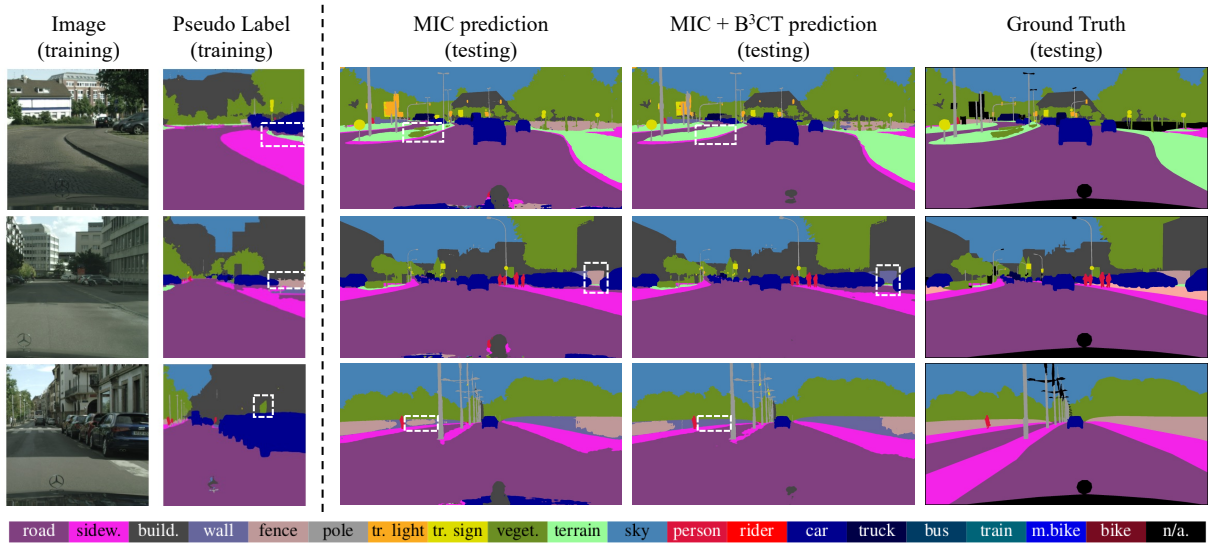


Fig. 6: **Failure cases of B^{3}CT on the Cityscapes validation set.** The white dashed boxes highlight regions where B^{3}CT performs worse than the MIC baseline, particularly for thin or texture-ambiguous classes such as *fence* and *terrain*. B^{3}CT overfits to noisy pseudo-label structures and yields inferior predictions compared with the baseline.

weather conditions. These models tend to produce blended predictions across category boundaries, resulting in inaccurate segmentation of complex regions. In contrast, MIC + B^{3}CT produces clearer and more coherent boundaries, even in cases where pixel-level differences are

Table 7: **Comparison of runtime and parameters. (GTAV→Cityscapes)** Runtime and parameters of different methods on an Nvidia Titan RTX.

Method	Throughput(img/s)	TFLOPs	Training params.	mIoU
DAFormer	0.7	1.6	85.2M	60.9
DAFormer + B^{3}CT	0.7	1.6	85.6M	62.6
HRDA	0.8	3.3	85.7M	73.8
HRDA + B^{3}CT	0.8	3.3	86.1M	74.8
MIC	0.8	3.3	85.7M	75.9
MIC + B^{3}CT	0.8	3.3	86.1M	76.4

Table 8: **Training Overhead Comparison (MIC baseline on Titan RTX).**

Method	Training Time (s/step)	Peak GPU Memory (MB)	mIoU (GTAV→CS)
MIC (Baseline)	5.24	22,404	75.9%
MIC + B^{3}CT	8.51	19,334	76.4%

subtle but semantically meaningful. By leveraging attention-based alignment between domains, our framework stabilizes the learning of target representations—particularly for ambiguous regions. The incorporation of source-domain features through the hybrid-attention branch aids in refining pseudo labels and enhances feature discriminability. Overall, our approach demonstrates improved robustness and segmentation quality in diverse and visually complex environments.

Despite the overall improvements, Fig. 6 shows representative failure cases when using MIC as the backbone, where classes such as *fence* and *terrain* exhibit degraded performance compared with MIC alone. These cases typically correspond to regions where the target-domain pseudo labels are highly noisy. Because the hybrid branch, AAC, and coordinate weighting all rely on the pseudo-label distribution for supervising cross-domain alignment, erroneous pseudo labels can mislead the attention fusion process and cause the model to strengthen incorrect patterns. As a result, B^{3}CT may overfit to noisy supervision signals and produce errors not present in the baseline.

These observations highlight that the effectiveness of B^{3}CT is closely linked to pseudo-label reliability. While the good cases demonstrate the strong potential of our design in leveraging cross-domain cues, the failure cases emphasize the need for more robust confidence modeling or noise-aware regularization. Exploring such mechanisms represents a promising direction for future extensions of B^{3}CT.

4.6 Comparison of computational efficiency

The runtime and parameter statistics of our framework during inference are reported in Table 7. A key advantage of our approach is that the hybrid branch and the AAC module are exclusively utilized during training and are completely removed at inference time. As a result, B3CT incurs no additional overhead during testing, maintaining identical throughput and TFLOPs to the baseline models.

To provide a complete picture, we also analyzed the training-time cost, as this is important for practical applications. While our method introduces an additional forward and backward pass for the hybrid branch, leading to a predictable increase in training duration, it offers a notable advantage in memory efficiency. As shown in Table 8, we compare the training overhead using MIC as the baseline. The training time per step increases from 5.24 s/step to 8.51 s/step. However, by enabling the effective and stable use of mixed-precision training, the peak GPU memory footprint of MIC + B3CT is reduced to 19,334 MB from the baseline's 22,404 MB.

In summary, B^{3}CT achieves superior segmentation performance at the cost of a moderate

increase in training time, but with the added benefit of lower memory consumption. This makes our method not only effective but also practical for training on standard 24GB GPUs, especially considering its zero-cost inference.

5 Conclusion

In this work, we address the two fundamental challenges in domain adaptive semantic segmentation: "where to align" and "when to align". To this end, we propose a Three-Branch Coordinated Training (B^3CT) framework, where the third branch integrates intra-domain self-attention and inter-domain cross-attention to facilitate effective feature fusion and alignment. An Adaptive Alignment Controller (AAC) is introduced to selectively perform alignment at appropriate network stages, while a coordinate weight mechanism is proposed to dynamically regulate the timing of alignment during training. Our method effectively balances learning discriminative category features from the source domain and adapting to the feature distribution of the target domain. Extensive experiments demonstrate that B^3CT achieves competitive performance on both GTAV→Cityscapes and SYNTHIA→Cityscapes benchmarks.

6 Limitation and Discussion

We acknowledge several limitations of our current work and outline directions for future research. First, the present framework is specifically designed for semantic segmentation tasks, with alignment targeted at the patch level; generalizing this approach to other vision tasks requires further exploration of architecture and training adaptations.

Second, while our method is developed for accessible target domain data during training, future work could investigate its extension to domain generalization settings and integration with vision foundation models for broader applicability.

Third, the field is rapidly moving towards more diverse alignment challenges. For example, emerging datasets like Ego-Exo4D [67] introduce instance-level cross-view segmentation tasks, where the goal is to transfer a segmentation mask from one viewpoint to another in a "one-shot" manner. Adapting the principles of our hybrid-attention mechanism to tackle such instance-specific correspondence problems, potentially in a training-free or few-shot context, presents a compelling and non-trivial direction for future research.

Finally, the performance of our method can be sensitive to certain hyperparameters, particularly in the coordination module. Although we found effective configurations empirically, developing more adaptive strategies that are robust across different datasets and model architectures remains an area for future improvement.

Declarations

- **Availability of data and materials.** All data will be made available on reasonable request.
- **Conflict of interest.** All authors declare no conflicts of interest.
- **Code availability.** The toolkit and experimental results will be made publicly available.

References

[1] Zhang, Y., Qiu, Z., Yao, T., Liu, D., Mei, T.: Fully convolutional adaptation networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6810–6818 (2018)

[2] Pan, Y., Yao, T., Li, Y., Wang, Y., Ngo, C.-W., Mei, T.: Transferrable prototypical networks for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2239–2247 (2019)

[3] Yang, J., An, W., Wang, S., Zhu, X., Yan, C., Huang, J.: Label-driven reconstruction for domain adaptation in semantic segmentation. In: European Conference on Computer Vision, pp. 480–498 (2020). Springer

[4] Kim, M., Byun, H.: 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)

[5] Cheng, Y., Wei, F., Bao, J., Chen, D., Wen, F., Zhang, W.: Dual path learning for domain adaptation of semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9082–9091 (2021)

[6] Shen, F., Gurram, A., Liu, Z., Wang, H., Knoll, A.: Diga: Distil to generalize and then adapt for domain adaptive semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15866–15877 (2023)

[7] Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)

[8] Zhang, P., Zhang, B., Zhang, T., Chen, D., Wang, Y., Wen, F.: 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)

[9] Liu, Y., Deng, J., Gao, X., Li, W., Duan, L.: 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 (2021)

[10] Wang, W., Zhong, Z., Wang, W., Chen, X., Ling, C., Wang, B., Sebe, N.: Dynamically instance-guided adaptation: A backward-free approach for test-time domain adaptive semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 24090–24099 (2023)

[11] Das, A., Xian, Y., Dai, D., Schiele, B.: Weakly-supervised domain adaptive semantic segmentation with prototypical contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15434–15443 (2023)

[12] Zhou, C., Wang, Z., Zhang, X., Du, B.: Domain complementary adaptation by leveraging diversity and discriminability from multiple sources. IEEE Transactions on Multimedia (2023)

[13] Liu, Y., Wang, J., Wang, W., Hu, Y., Wang, Y., Xu, Y.: Crada: Cross domain object detection with cyclic reconstruction and decoupling adaptation. IEEE Transactions on Multimedia (2024)

[14] Tsai, Y.-H., Hung, W.-C., Schulter, S., Sohn, K., Yang, M.-H., Chandraker, M.: 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)

[15] Vu, T.-H., Jain, H., Bucher, M., Cord, M., Pérez, P.: 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)

[16] Gong, R., Wang, Q., Danelljan, M., Dai, D., Van Gool, L.: Continuous pseudo-label rectified domain adaptive semantic segmentation with implicit neural representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7225–7235 (2023)

[17] Li, C., Zhang, D., Huang, W., Zhang, J.: Cross contrasting feature perturbation for domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1327–1337 (2023)

[18] Ge, F., Zhang, Y., Wang, L., Coleman, S., Kerr, D.: Double-domain adaptation semantics for retrieval-based long-term visual localization. IEEE Transactions on Multimedia (2023)

[19] Chen, T., Wang, S.-H., Wang, Q., Zhang, Z., Xie, G.-S., Tang, Z.: Enhanced feature alignment for unsupervised domain adaptation of semantic segmentation. IEEE Transactions on Multimedia **24**, 1042–1054 (2021)

[20] Fan, B., Yang, Y., Feng, W., Wu, F., Lu, J., Liu, H.: Seeing through darkness: Visual localization at night via weakly supervised learning of domain invariant features. IEEE Transactions on Multimedia (2022)

[21] Yang, J., An, W., Yan, C., Zhao, P., Huang, J.: Context-aware domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 514–524 (2021)

[22] Rao, P.P., Qiao, F., Zhang, W., Xu, Y., Deng, Y., Wu, G., Zhang, Q.: Quadformer: Quadruple transformer for unsupervised domain adaptation in power line segmentation of aerial images. arXiv preprint arXiv:2211.16988 (2022)

[23] Wang, X., Guo, P., Zhang, Y.: Domain adaptation via bidirectional cross-attention transformer. arXiv preprint arXiv:2201.05887 (2022)

[24] Xu, T., Chen, W., Wang, P., Wang, F., Li, H., Jin, R.: Cdtrans: Cross-domain transformer for unsupervised domain adaptation. arXiv preprint arXiv:2109.06165 (2021)

[25] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)

[26] Yuan, Y., Huang, L., Guo, J., Zhang, C., Chen, X., Wang, J.: Ocnet: Object context network for scene parsing. arXiv preprint arXiv:1809.00916 (2018)

[27] Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16, pp. 173–190 (2020). Springer

[28] Poudel, R.P., Bonde, U., Liwicki, S., Zach, C.: Contextnet: Exploring context and detail for semantic segmentation in real-time. arXiv preprint arXiv:1805.04554 (2018)

[29] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)

[30] Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K.: Denseaspp for semantic segmentation in street scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3684–3692 (2018)

[31] Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3146–3154 (2019)

[32] Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)

[33] Zhao, H., Zhang, Y., Liu, S., Shi, J., Loy, C.C., Lin, D., Jia, J.: Psanet: Point-wise spatial attention network for scene parsing. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 267–283 (2018)

[34] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[35] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)

[36] Li, Y., Yuan, L., Vasconcelos, N.: 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)

[37] Kumar, A., Sattigeri, P., Wadhawan, K., Karlinsky, L., Feris, R., Freeman, B., Wornell, G.: Co-regularized alignment for unsupervised domain adaptation. Advances in neural information processing systems **31** (2018)

[38] Zou, Y., Yu, Z., Kumar, B., Wang, J.: 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)

[39] Zou, Y., Yu, Z., Liu, X., Kumar, B., Wang, J.: Confidence regularized self-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5982–5991 (2019)

[40] Mei, K., Zhu, C., Zou, J., Zhang, S.: Instance adaptive self-training for unsupervised domain adaptation. In: European Conference on Computer Vision, pp. 415–430 (2020). Springer

[41] Wang, Y., Peng, J., Zhang, Z.: Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9092–9101 (2021)

[42] Araslanov, N., Roth, S.: 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)

[43] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[44] Chen, C.-F.R., Fan, Q., Panda, R.: 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)

[45] Tang, Z., Qiu, Z., Hao, Y., Hong, R., Yao, T.: 3d human pose estimation with spatio-temporal criss-cross attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4790–4799 (2023)

[46] Kang, X., Wang, C., Chen, X.: Region-enhanced feature learning for scene semantic segmentation. IEEE Transactions on Multimedia, 1–11 (2023) https://doi.org/10.1109/TMM.2023.3342718

[47] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021)

[48] Hoyer, L., Dai, D., Van Gool, L.: 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)

[49] Hoyer, L., Dai, D., Van Gool, L.: Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In: European Conference on Computer Vision, pp. 372–391 (2022). Springer

[50] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., *et al.*: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)

[51] Dai, Y., Liu, J., Sun, Y., Tong, Z., Zhang, C., Duan, L.-Y.: Idm: An intermediate domain module for domain adaptive person re-id. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11864–11874 (2021)

[52] Deng, J., Zhang, X., Li, W., Duan, L., Xu, D.: Cross-domain detection transformer based

on spatial-aware and semantic-aware token alignment. IEEE Transactions on Multimedia (2023)

[53] Zhang, W., Ouyang, W., Li, W., Xu, D.: Collaborative and adversarial network for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

[54] Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)

[55] Huang, J., Guan, D., Xiao, A., Lu, S.: Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. Advances in Neural Information Processing Systems **34**, 3635–3649 (2021)

[56] Luo, Y., Zheng, L., Guan, T., Yu, J., Yang, Y.: Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2507–2516 (2019)

[57] Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, pp. 102–118 (2016). Springer

[58] Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: 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)

[59] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)

[60] Tranheden, W., Olsson, V., Pinto, J., Svensson, L.: 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)

[61] Wang, Q., Dai, D., Hoyer, L., Van Gool, L., Fink, O.: Domain adaptive semantic segmentation with self-supervised depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8515–8525 (2021)

[62] Hoyer, L., Dai, D., Wang, H., Van Gool, L.: Mic: Masked image consistency for context-enhanced domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11721–11732 (2023)

[63] Huo, X., Xie, L., Zhou, W., Li, H., Tian, Q.: Focus on your target: A dual teacher-student framework for domain-adaptive semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 19027–19038 (2023)

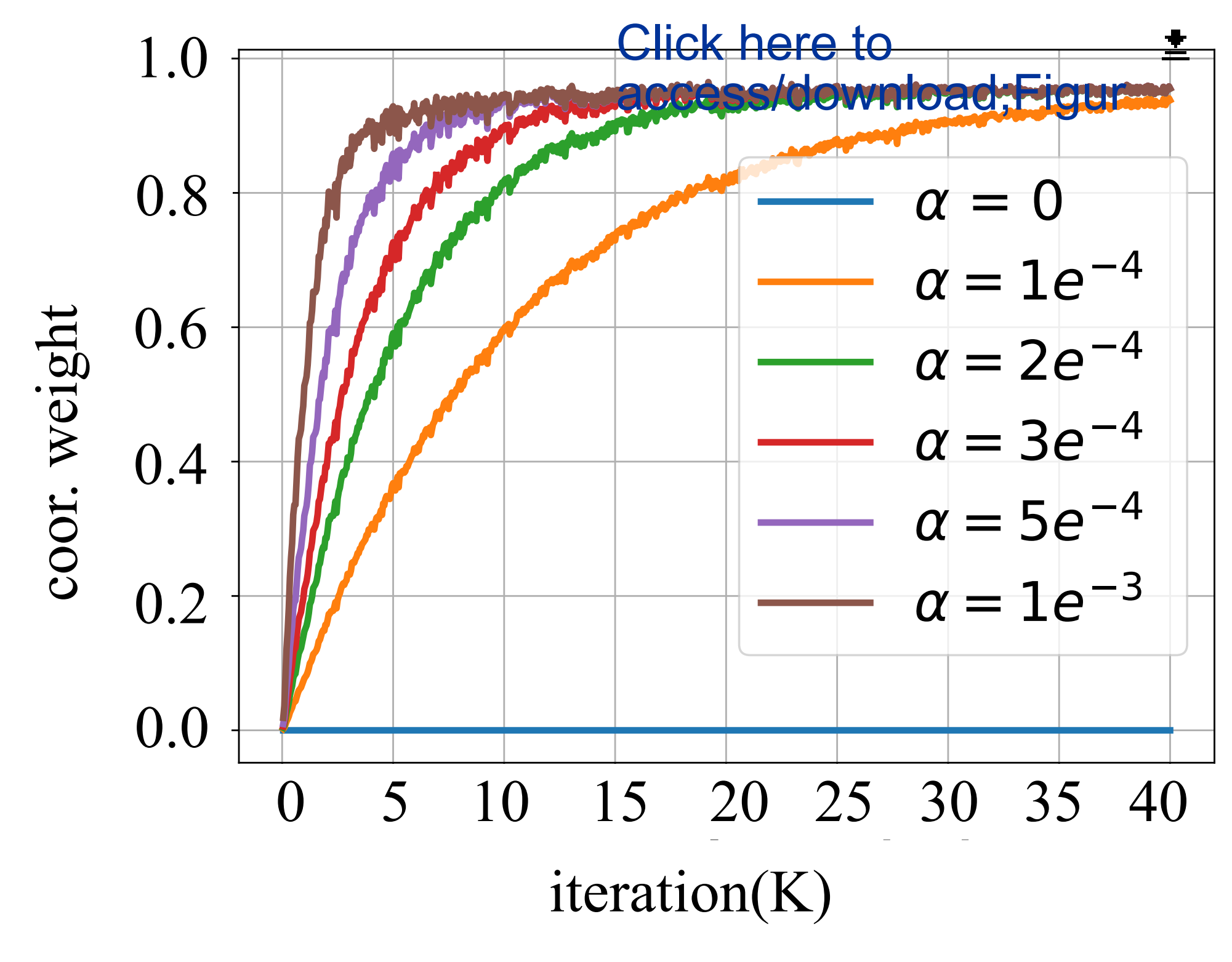
[64] Sakaridis, C., Dai, D., Van Gool, L.: Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(6), 3139–3153 (2020)

[65] Wu, X., Wu, Z., Guo, H., Ju, L., Wang, S.: Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15769–15778 (2021)

[66] Zheng, Z., Yang, Y.: Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. International Journal of Computer Vision **129**(4), 1106–1120 (2021)

[67] Fu, Y., Wang, R., Fu, Y., Paudel, D.P.,

Van Gool, L.: Cross-view multi-modal segmentation@ ego-exo4d challenges 2025. arXiv preprint arXiv:2506.05856 (2025)


Click here to access/download;Figur
1.0
0.8
0.6
0.4
0.2
0.0
coor. weight
0
5
10
15
20
25
30
35
40
iteration(K)
$\alpha = 0$
$\alpha = 1e^{-4}$
$\alpha = 2e^{-4}$
$\alpha = 3e^{-4}$
$\alpha = 5e^{-4}$
$\alpha = 1e^{-3}$