Supplemental Material
SUMMIT: Source-Free Adaptation of Uni-Modal Models to Multi-Modal Targets

Cody Simons[1] Dripta S. Raychaudhuri[1,2,*] Sk Miraj Ahmed[1] Suya You[3]
Konstantinos Karydis[1] Amit K. Roy-Chowdhury[1]
[1]University of California, Riverside [2]AWS AI Labs [3]DEVCOM Army Research Laboratory
`{csimo005@,drayc001@,sahme047@,kkarydis@ece.,amitrc@ece.}ucr.edu` `suya.you.civ@army.mil`

A. Training Pseudo-Code

We present the pseudo-code for our training algorithm in Algorithm 1. The mathematical notation is the same as that described in Section 3.1 of the main paper.

B. Additional training details

Table A-1: Here we show the hyper parameters for each experiment. The hyperparameters are the same for USA/Singapore & the Day/Night experiments since we expect a similar domain gap. We modify the learning rate and λ_{xM} slight for the A2D2/SemanticKITTI and all crossover experiments, since they have a larger domamin gap.

	USA/Singapore	Day/Night	A2D2/SemanticKITTI	Crossover
Optimizer	Adam	Adam	Adam	Adam
Learning Rate	1e-5	1e-5	1e-3	1e-3
β_1	0.9	0.9	0.9	0.9
β_2	0.999	0.999	0.999	0.999
Scheduler	MultiStep	MultiStep	MultiStep	MultiStep
Learning Rate Decay	0.1	0.1	0.1	0.1
Milestone	80K, 90K	80K, 90K	80K, 90K	80K, 90K
Max Iteration	100K	100K	100K	100K
Batch Size	8	8	8	8
λ_{xM} Target	0.1	0.1	0.01	0.01

Most of our hyperparameters we take directly from [1], however we reduce the initial learning rate on the USA/Singapore and the Day/Night adaptation scenario. This is done because of the smaller domain gaps in these two adaptation scenarios. The initial learning rate in the case of the A2D2/SemanticKITTI adaptation scenario is relatively larger because of the larger domain gap. The value of λ_{xM} is set according to [1], including the lower value being used for the A2D2/SemanticKITTI adaptation scenario.

C. Hypothesis Testing Threshold Analysis

In this section we present a sensitivity analysis of the hypothesis testing portion of entropy weighting to the threshold value. We show the results for thresholds of 0.5, 1, and 2, which correspond to switching if the alternative is at least half as likely, just as likely, and twice as likely as the alternative. Unsurprisingly when $\tau = 0.5$, we see the lowest performance, since in this case we chose the alternative when it's less likely than the null hypothesis. However, in this case the performance is still on par with the unadapted 2D and better than the unadapted 3D, which shows that even with a poorly tuned threshold the method does not hurt performance. In the case where

Algorithm 1 SUMMIT

```
Require: : Uni-Modal Source Models - $\mathcal{M}^{2D}$ & $\mathcal{M}^{3D}$, Source Model Metrics - $Top1^{2D}$ & $Top1^{3D}$,Multi-Modal Target Dataset - $\mathcal{D}_\mathcal{T}$
for $\{x_i^{2D}, x_i^{3D}\} \in \mathcal{D}_\mathcal{T}$ do
    $\tilde{y}_i^{2D} = \begin{cases} \arg\max_k \mathcal{M}_k^{2D}(x_i^{2D}), & \mathcal{M}_k^{2D}(x_i^{2D}) \geq \text{median}_k^{2D} \\ \text{ignore}, & otherwise \end{cases}$
    $\tilde{y}_i^{3D} = \begin{cases} \arg\max_k \mathcal{M}_k^{3D}(x_i^{3D}), & \mathcal{M}_k^{3D}(x_i^{2D}) \geq \text{median}_k^{3D} \\ \text{ignore}, & otherwise \end{cases}$
end for
Source Agreement= $Top1^{2D} \cdot Top1^{3D}$
Target Agreement= $\frac{\sum_{\tilde{y}^{2D}, \tilde{y}^{3D}} \mathbb{1}(\tilde{y}^{2D} == \tilde{y}^{3D})}{|\mathcal{D}_\mathcal{T}|}$
if $\frac{\text{Source Agreement}}{\text{Target Agreement}} \leq 0.5$ then
    for $\tilde{y}_i^{2D}$ & $\tilde{y}_i^{3D}$ do
        $\tilde{y}_i = \begin{cases} \tilde{y}_i^{2D}, & \tilde{y}_i^{2D} = \tilde{y}_i^{3D} \\ \text{ignore}, & \tilde{y}_i^{2D} \neq \tilde{y}_i^{3D} \end{cases}$
    end for
else
    for $\{x_i^{2D}, x_i^{3D}\} \in \mathcal{D}_\mathcal{T}$ do
        $w^{2D} = \frac{e^{-h(\mathcal{M}^{2D}(x^{2D}))}}{e^{-h(\mathcal{M}^{2D}(x^{2D}))} + e^{-h(\mathcal{M}^{3D}(x^{3D}))}}$
        $w^{3D} = 1 - w^{2D}$
        $p_i = w^{2D}\psi(\mathcal{M}^{2D}(x^{2D})) + w^{3D}\psi(\mathcal{M}^{3D}(x^{3D}))$
        $\tilde{y}_i = \begin{cases} \arg\max_k p_i, & p_{i,k} \geq \text{median}_k^p \\ \text{ignore}, & otherwise \end{cases}$
        if $\tilde{y}_i$ is ignored then
            $R_{2D} = \frac{\mathcal{N}(f^{2D}(x^{2D}); \mu_{k_{2D}}^{2D}, \left(\sigma_{k_{2D}}^{2D}\right)^2)}{\mathcal{N}(f^{2D}(x^{2D}); \mu_{k_{3D}}^{2D}, \left(\sigma_{k_{3D}}^{2D}\right)^2)}$
            $R_{3D} = \frac{\mathcal{N}(f^{3D}(x^{3D}); \mu_{k_{3D}}^{3D}, \left(\sigma_{k_{3D}}^{3D}\right)^2)}{\mathcal{N}(f^{3D}(x^{3D}); \mu_{k_{2D}}^{3D}, \left(\sigma_{k_{2D}}^{3D}\right)^2)}$
            if $R_{2D} \leq \tau$ and $R_{3D} > \tau$ then
                $\tilde{y}_i = \tilde{y}_i^{3D}$
            end if
            if $R_{2D} > \tau$ and $R_{3D} \leq \tau$ then
                $\tilde{y}_i = \tilde{y}_i^{2D}$
            end if
        end if
    end for
end if
for $K$ iterations do
    Sample $\{x^{2D}, x^{3D}, \tilde{y}\}$ from $\mathcal{D}_\mathcal{T}$
    Calculate $\mathcal{L}_{tot}(\mathcal{M}^{2D}(x_i^{2D}), \mathcal{M}^{3D}(x_i^{3D}), \tilde{y})$
    Update $\mathcal{M}^{2D}, \mathcal{M}^{3D}$ to minimize $\mathcal{L}_{tot}$
end for
```

$\tau = 2$, we see that there are still minor improvements across the board. We see strongest improvement when $\tau = 1$, which corresponds to only switching when one is more likely than the other.

Table A-2: Here we present a analysis of the sensitivity of the hypothesis testing threshold. We note that while there is some variation in performance, our performance is still on par with unadapted performance. So while a better tuned value may help performance, poorly tuned values will not hurt performance.

	USA/Singapore		
τ	2D	3D	2D+3D
0.5	49.25	50.18	53.39
1	57.47	52.12	62.32
2	50.27	47.97	56.19

D. Crossover Automatic Switching Analysis

We include here the full analysis of the pseudo-label switching method for the cross-over experiments. We present the full results in Table A-3. We can see that the agreement filtering is indeed selected throughout the crossover experiments. We note the lower source agreement compared to the first three experiments. This is likely due to the Singapore & Night splits of NuScenes having fewer samples to train the source model, about 10K and 3K respectively, compared to the USA & Day splits which have 16K and 25K. However, since the target agreement is much lower the ratio between them still stays below our threshold of 0.5, so AF is selected.

The accuracy of the filtered pseudo-labels is presented in Table A-4. Once again we see that AF admits far fewer pseudo-labels, but the vast majority of these are correct. If we look at entropy weighting we see that in this case it has failed quite dramatically, admitting mostly incorrect labels. EW performs so poorly here because of the low source model performance, as can be seen in the no adaptation row of Table 6 in the main paper.

Table A-3: In crossover experiments we see that the source agreement is lower across the board. However, because of the domain gap between the source and target we see that the target agreement is far lower as well. This results in the crossover experiments consistently using the agreement filtering method.

	USA 2D, Sing. 3D	**Sing. 2D, USA 3D**	**Day 2D, Night 3D**	**Night 2D, Day 3D**
Source Agreement	86.37	85.23	81.24	75.09
Target Agreement	38.60	30.60	32.80	35.00
Ratio	0.45	0.36	0.40	0.47

Table A-4: In crossover experiments we see that the agreement filtering correctly labels a much larger portion of admitted samples. In contrast, the statistical fusion method admits many samples, but the pseudo-labels are generally incorrect. This indicates a larger portion of noise hurting the statistical fusion process.

	(USA-2D,Singapore-3D)/SemKITTI			**(Singapore-2D,USA-3D)/SemKITTI**			**(Day-2D,Night-3D)/SemKITTI**			**(Night-2D,Day-3D)/SemKITTI**		
Method	Correct	Incorrect	Ignore	Correct	Incorrect	Ignore	Correct	Incorrect	Ignore	Correct	Incorrect	Ignore
AF	28.66	3.23	68.11	30.78	3.38	65.84	32.76	4.77	62.46	26.44	3.19	70.37
EW	7.49	85.28	7.23	4.26	89.92	5.82	4.41	89.97	5.62	1.41	94.88	3.71

E. Description of Attached Video

We have included two videos named `A2D2SemanticKITTI.mp4` and `NuscenesLidarsegSemanticKITTI.mp4`. Both videos show a clip from the SemanticKITTI dataset with the different colored points corresponding to the label predictions of a model adapted using agreement filtering. In `A2D2SemanticKITTI.mp4` the source models are trained on the A2D2 dataset and in `NuscenesLidarsegSemanticKITTI.mp4` the 2D model is trained using the Day split of NuScenes and the 3D model is trained using the Night split, corresponding to the *crossover* experiments. In both videos we show the predicted label for 2D & 3D individually, the combined 2D+3D, and the ground truth classification. Please note that each frame is evaluated individually, making no use of any temporal correlations which we leave to future works.

References

[1] Jaritz, M., Vu, T.H., Charette, R.d., Wirbel, E., Pérez, P.: xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In: CVPR (2020) A-1