Continuous Pseudo-Label Rectified Domain Adaptive Semantic Segmentation with Implicit Neural Representations
— Supplementary Material —

Rui Gong[1] Qin Wang[1] Martin Danelljan[1] Dengxin Dai[2] Luc Van Gool[1]
[1]CVL, ETH Zurich [2]Max Planck Institute for Informatics, Saarland Informatics Campus
{gongr, qin.wang, martin.danelljan, vangool}@vision.ee.ethz.ch, ddai@mpi-inf.mpg.de

In this supplementary material, we provide the additional information for,

S1 code availability, and discussion about limitations,

S2 details of our implementation and experimental settings,

S3 details of the datasets involved in our experiments,

S4 experimental results on generalization ability measurement of the trained model,

S5 comparisons to the structure variant of plugging continuous RMM into HRDA,

S6 experimental results of predicting rectification values on unseen coordinates,

S7 discussion on motivation and structure design,

S8 more quantitative and qualitative experimental results.

S1. Code Availability and Limitations Analysis

Code Availability. Our implementation is available at `https://github.com/ETHRuiGong/IR2F`.

Limitations. In this paper, we propose the continuous rectification-aware mixture model (RMM) based on implicit rectification-representative function (IR^{2}F), *i.e.* continuous RMM. We analyze limitations of our approach from two aspects: 1) Though our continuous RMM is proven effective for rectifying pseudo-labels and boosting the UDA performance by a large margin, our continuous RMM is still yet to achieve the performance of fully supervised learning. 2) When testing our adapted model with another unseen dataset, *i.e.* domain generalization scenario, we can observe that our generalization performance falls behind our achieved UDA performance. This is the same as other UDA methods. However, as shown in Table S1, our continuous RMM still strongly outperforms the previous SOTA UDA methods under the generalization scenario. Even though the further investigation on, how to improve the generalization performance of the UDA model, is out of the scope of this work, we believe that this is an interesting aspect for future work, unifying the domain adaptation and generalization scenario and learning more adaptive and generalizable models.

S2. Implementation and Training Details

In the main paper, we propose the IR^{2}F-based continuous RMM for UDA, which can be used as a plug-in module and is compatible with different UDA frameworks. In Sec. 4.1 of the main paper, we provide implementation details of the framework structure and training. Here we present additional more detailed implementation of our proposed continuous RMM.

Training Details. By default, we follow the training details of HRDA [7]. We utilize AdamW [9, 11] optimizer, where betas of AdamW optimizer are (0.9, 0.999), the weight decay is 0.01, and learning rates of the encoder and decoder are set as 6×10^{-5}, 6×10^{-4}, respectively. The batch size is set as 2, and the linear learning rate warmup and DACS [19] data augmentation in [6, 7] are adopted. For Table 4 of the main paper (our method integrated with MRNet), we follow the training details of MRNet [27].

IR^{2}F. $f_{\theta'}$ is implemented with a 4-layer multi-layer perceptron (MLP), with the hidden dimension of 256 and ReLU activation.

S3. Datasets Information

As introduced in Sec. 4 of the main paper and Sec. S4 of the supplementary, there are 6 datasets involved in our experiments, including GTA [14], SYNTHIA [15], Cityscapes [4], Dark Zurich [16], ACDC-Night [17] and NightCity+ [5, 18]. In this section, we provide the detailed information about these datasets.

GTA. GTA [14] is a synthetic urban-scene image dataset, rendered from the game engine. There are 24966 images included in the GTA dataset, which are of 1914×1052 pixels and are densely labeled with pixel-wise semantic segmentation annotations. The urban scene of GTA dataset is built based on the city of Los Angeles, thus with typical U.S. urban scene layout. Following previous UDA works [6,7,19–21,25,27,28], the GTA images are resized to 1280×720 for low-resolution inputs [7], and to 2560×1440 for high-resolution inputs [7].

SYNTHIA. SYNTHIA [15] is a synthetic photo-realistic image dataset, whose images are rendered from a virtual city. We adopt SYNTHIA-RAND-Cityscapes dataset, which is built for street scene parsing and consists of 9400 densely labeled images. The images are of 1280×760 pixels. In accordance with previous UDA works [6, 7, 19–21, 25,27,28], the SYNTHIA images are resized to 2560×1520 for high-resolution inputs [7], and keep 1280×760 for low-resolution inputs.

Cityscapes. Cityscapes [4] is a real street-scene image dataset, collected from different European cities. We utilize the training set of Cityscapes during the training stage, consisting of 2975 images. And we use the validation set of Cityscapes, covering 500 images, to evaluate the model performance. Cityscapes images are of 2048×1024 pixels. The resolution is maintained for high-resolution inputs in experiments, and resized to 1024×512 for low-resolution inputs.

Dark Zurich. Dark Zurich [16] is a real nighttime urban-scene image dataset, which is captured in Zurich. We use the training set of Dark Zurich during the training stage, including 2416 images. And we utilize the test set of Dark Zurich, consisting of 151 images, to evaluate the model performance. The evaluation on the test set of Dark Zurich is only accessible through the online benchmark, where the ground truth is not publicly available. The images in Dark Zurich is of 1920×1080 pixels. The resolution is kept for high-resolution inputs, and is resized to 960×540 for low-resolution inputs.

ACDC-Night. ACDC [17] is a real street-scene image dataset under adverse conditions, *e.g.* fog, snow, rain and nighttime. We adopt the nighttime subset of ACDC, *i.e.* ACDC-Night, where there are 400 images as training set and 500 images as test set. Similar to the evaluation of Dark Zurich, the evaluation on the test set can only be conducted through the online benchmark, and the ground truth is not publicly available. The images in ACDC-Night is of 1920×1080 pixels. The resolution is kept for high-resolution inputs, and is resized to 960×540 for low-resolution inputs.

NightCity+. NightCity [18] is a real urban driving scene dataset, for nighttime scene parsing. The images in NightCity are collected from different cities around the world. NightCity+ [5] is the extended version of NightCity, where more accurate annotations in the validation set are provided compared to NightCity. We utilize the validation set of NightCity+, including 1299 images, to evaluate the model performance.

Method	PSPNet [26]	DANNet [22]	DANIA [23]	GLASS [10]	HRDA [7]	Ours
mIoU (%)	19.0	29.9	28.9	31.8	36.7	**38.5**

Table S1. **Quantitative generalization comparisons**, on NightCity+ dataset. The model is trained on the day-to-night benchmark, Cityscapes→Dark Zurich, and is tested on NightCity+ dataset.

S4. Generalization Experiments

In Sec. 4 of the main paper, we compare our continuous RMM to other methods under the UDA setting, proving the advantage of our approach for domain adaptation. Following [10], we further evaluate the adapted model (after domain adaptation) performance on another unseen dataset, to show the generalization ability of our proposed continuous RMM. More specifically, we take the trained model under Cityscapes→Dark Zurich (*i.e.* the model in Table 2 of the main paper), and evaluate the trained model on the third dataset NightCity+.

Quantitative Experimental Results. As shown in Table S1, our proposed approach strongly outperforms other UDA methods on the generalization ability evaluation, 38.5% *vs.* 36.7%, 31.8%, 28.9%, 29.9%, 19.0%. It proves that our model trained on Cityscapes→Dark Zurich generalizes well to other unseen nighttime datasets.

Qualitative Experimental Results. In Fig. S1, we show the qualitative comparisons between our proposed approach and previous SOTA method, HRDA [7], for generalization. It is further proven that our trained model can generalize well to other unseen nighttime datasets.

S5. Structure Variant Comparisons

Our proposed continuous RMM can be used as a plug-in strategy to promote and rectify pseudo-labels used for self-training in UDA. We introduced the structure of plugging continuous RMM into HRDA in Sec. 3.4 and Fig. 3 of the main paper. Here we compare to the structure variant of plugging continuous RMM. As shown in Fig. S2, besides the structure in the main paper, we propose another structure variant by training an IR^{2}F for each branch. The estimated rectification values $\mathbf{r}_1, \mathbf{r}_2$ are normalized as $\frac{\mathbf{r}_1}{\mathbf{r}_1+\mathbf{r}_2}, \frac{\mathbf{r}_2}{\mathbf{r}_1+\mathbf{r}_2}$, to satisfy $\mathbf{r}_1 + \mathbf{r}_2 = 1$. Then we quantitatively compare the performance of the structure in the main paper to this structure variant, under the SYNTHIA→Cityscapes benchmark. As shown in Table S2, the structure in the main paper and this variant achieve very similar performance 67.7% *vs.* 67.6% on the benchmark, SYNTHIA→Cityscapes. It further proves the effectiveness of our proposed IR^{2}F based continuous RMM for UDA, and the flexibility of plugging

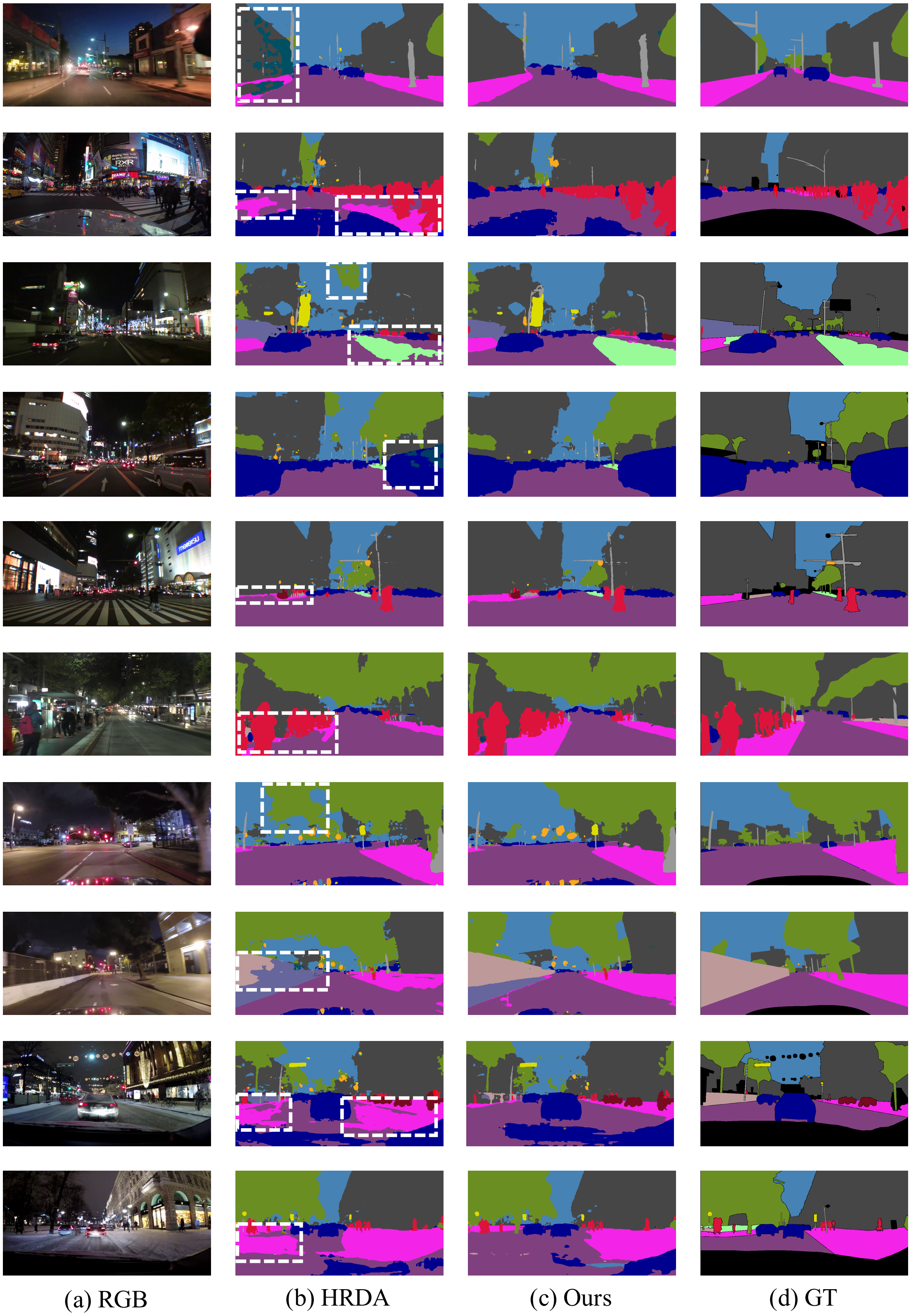


Figure S1. **Qualitative generalization comparisons**, on NightCity+ dataset. The model is trained on the day-to-night benchmark, Cityscapes→Dark Zurich, and is tested on NightCity+ dataset.

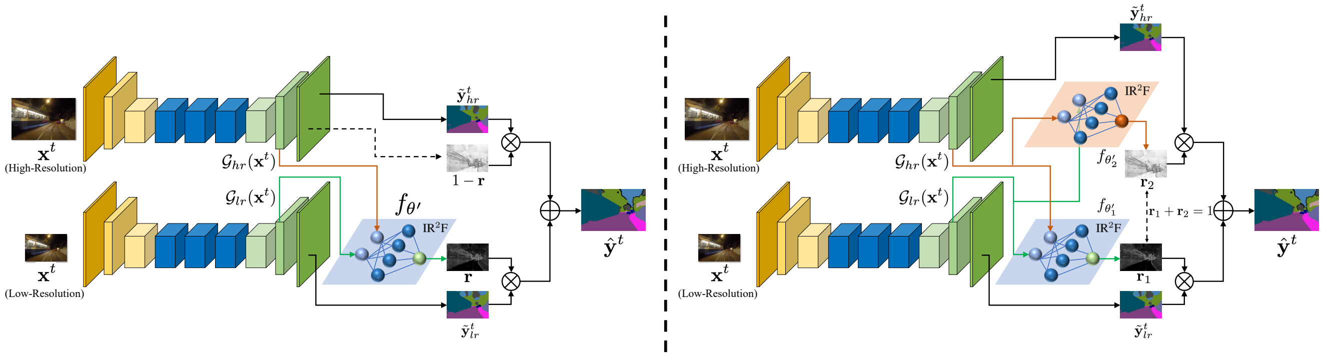


Figure S2. **Plugging continuous RMM into HRDA**, with different structures. The structure on the *left* is the structure that we adopt in the main paper. The structure on the *right* is the structure variant, where there is an IR^{2}F trained for each branch, and the estimated rectification values $\mathbf{r}_1, \mathbf{r}_2$ are normalized as $\frac{\mathbf{r}_1}{\mathbf{r}_1+\mathbf{r}_2}, \frac{\mathbf{r}_2}{\mathbf{r}_1+\mathbf{r}_2}$ to satisfy $\mathbf{r}_1 + \mathbf{r}_2 = 1$.

Method	Road	SW	Build	Wall	Fence	Pole	TL	TS	Veg	Sky	Person	Rider	Car	Bus	MC	Bike	mIoU
Original	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
Variant	90.0	53.8	89.3	46.9	7.2	59.7	65.0	63.6	87.6	93.4	80.3	54.4	89.9	69.1	65.4	65.9	67.6

Table S2. **Comparison to structure variant**, under SYNTHIA → Cityscapes. "Original" is the structure adopted in the main paper, and "Variant" is the structure which trains an IR^{2}F for each branch (see right side of Fig. S2). These two structures achieve similar performance under SYNTHIA→Cityscapes.

our continuous RMM into the UDA frameworks through different strategies.

S6. Rectification Values Prediction on Unseen Coordinates

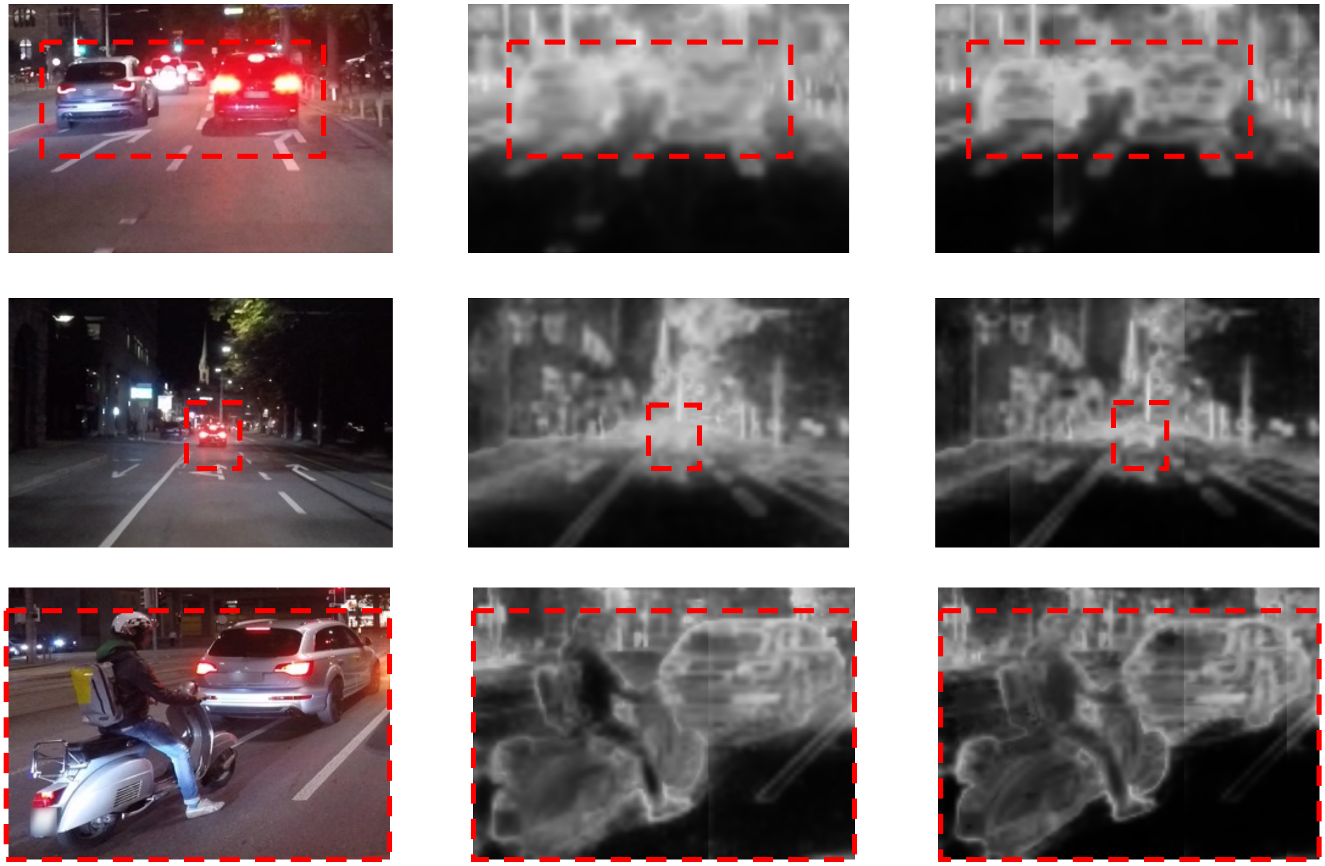
In order to test the generalization ability of rectification values prediction method to the unseen coordinates, we conduct the experiments to predict the rectification values on the 2× image coordinates. Since only the image coordinates are utilized during the training, the 2× image coordinates are unseen for training and are only used for testing. For the discrete modeling method HRDA [7], the prediction on unseen 2× image coordinates is realized by first predicting on the original image coordinate and then up-sampling (*e.g.* bilinear sampling) to the 2× image coordinates. For our continuous modeling method, IR^{2}F can directly output the rectification values by inputting 2× image coordinates. As shown in Fig. S3, it is observed that our IR^{2}F-based continuous modeling method generalizes well to the unseen coordinates and preserves finer details compared to the discrete modeling method, especially the boundary parts (see Fig. S3).

S7. Motivation and Structure Design Discussion

Where and why the continuous modelling is useful. As shown in Table 6 of the main paper, our IR^{2}F outperforms different baselines, including discrete convolutional decoder based rectification value estimation, 67.7% *vs*. 65.8%. This largely stems from the continuous modelling allowed by the implicit neural representations (INR). The continuous nature of INR is critical because *off-grid positions* exist in: 1) the feature's spatial dimension is smaller than the final output; 2) different mixture members in RMM can be of different resolutions. As shown in Fig. 5, our IR^{2}F achieves sharper and more accurate rectification values, while the discrete convolutional predictions are blurry and coarse due to sub-optimal interpolation. In Fig. S3 of the supplementary, compared to the discrete method, our IR^{2}F can generate higher quality predictions on higher resolutions that are unseen during training.

An INR for all images. Traditional INR works [1, 12] represent an object as a function, which maps coordinates to a signal (*e.g.* signed distance to a 3D object surface). However, instead of fitting individual functions for each object, more recent works [2, 3] aim to learn a shared general INR for multiple objects, to share knowledge across instances. The shared INR works typically utilize the encoder-based strategy, where different objects are mapped to different latent codes while sharing the same decoding INR function. The shared decoding function takes additional latent codes/ feature vector as input besides the coordinates. We follow the shared INR mechanism.

Relative coordinates *vs*. absolute coordinates. As shown in Eq. (2) of the main paper, in IR^{2}F, each local latent code from different mixture members is responsible for predicting rectification values of coordinates that are closest to it-



(a) RGB

(b) Discrete

(c) Continuous (**Ous**)

Figure S3. **Rectification values prediction on unseen coordinates.** Rectification values are predicted on the 2× image coordinates during the testing stage, which are unseen during the training stage. (b) is realized through the bilinear sampling of the output of HRDA [7], which is the discrete modeling method. (c) is realized by directly predicting rectification values on the 2× image coordinates with our IR^{2}F, which is continuous modeling method. It is shown that our IR^{2}F-based continuous modeling method can generalize well to the unseen 2× image coordinates, preserving finer details especially the boundary parts (see red dashed box).

self. Thus, the relative coordinate, instead of the absolute coordinate, between the queried coordinate and the coordinate of nearest local latent code, is input to IR^{2}F.

One feature *vs*. four nearest features. As shown in Eq. (2), IR^{2}F decodes relative coordinate (instead of distance) into the rectification value, conditioned on the nearest features. The difference between one and four nearest features is whether to incorporate more nearby features in the decoding process [2, 8, 24]. We conduct the experiment of four nearest features, yielding 74.5% mIoU for GTA→Cityscapes. The performance is similar to that of one nearest feature, 74.4%. To save memory, we take the one nearest feature strategy.

Performance gain origin. We compare our continuous method to other discrete learning-based methods, *e.g*. utilizing an additional convolutional decoder to predict rectification, in Table 6 and Table 5 of the main paper. Our approach outperforms these methods by at least +1.9 (67.7% *vs*. 65.8%), proving continuous nature's substantial advantage.

S8. More Quantitative and Qualitative Results

Oracle pseudo-label rectification accuracy. To explore the oracle pseudo-label rectification accuracy, we train HRDA+IR^{2}F in the supervised way, with cityscapes images and ground truth labels. Then under GTA→Cityscapes, mean pixel accuracy over all classes of our method and oracle are 81.65% *vs*. 88.28%.

Within-domain experiments. To study the effectiveness of our IR^{2}F for within-domain problems, we plug our continuous RMM module into CCT [13], as done in Sec. 3.4. For the semi-supervised segmentation task on PASCAL VOC (1/16 labeled images), our IR^{2}F+CCT improves CCT from 65.22% to 67.20%. It shows that our method helps for

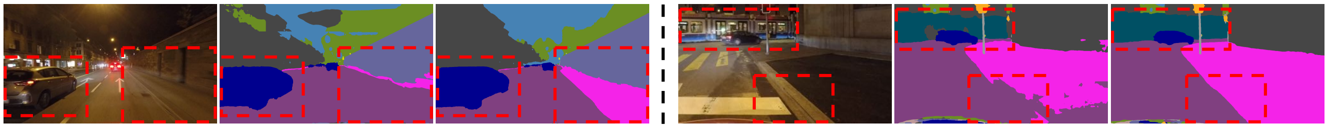


Figure S4. **Discrete *vs*. continuous pseudo-label visualizations.** In each group, RGB images (left), pseudo-label of discrete method (middle) and pseudo-label of our continuous method (right) are shown.

Learner Num	MRNet	MRNet+IR^{2}F	Improv.	Learner Num	MRNet	MRNet+IR^{2}F	Improv.
2	50.3	52.3	**+2.0**	3	51.1	52.9	**+1.8**

Table S3. **Learner Number Effects,** under GTA→Cityscapes.

within-domain problems.

Discrete *vs*. continuous pseudo-label visualizations. From the pseudo-label visualizations of discrete and continuous methods in Fig. S4, it is observed that our continuous method preserves finer details (*e.g*. sharp and accurate car, train and sidewalk boundaries) than the discrete method.

Increased number of learners. To explore the effectiveness of our continuous RMM when more learners are utilized, we compare MRNet+IR^{2}F performance of 2 and 3 learners, shown in Table S3. The improvement of 3 learners, brought by continuous RMM, is slightly lower than that of 2 learners, 1.8% *vs*. 2.0%. It proves that our continuous RMM method is still beneficial with increased number of learners.

References

[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, 2021. 4

[2] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *CVPR*, 2021. 4, 5

[3] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, 2019. 4

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1, 2

[5] Xueqing Deng, Peng Wang, Xiaochen Lian, and Shawn Newsam. NightLab: A dual-level architecture with hardness detection for segmentation at night. In *CVPR*, 2022. 1, 2

[6] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*, 2022. 1, 2

[7] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *ECCV*, 2022. 1, 2, 4, 5

[8] Hanzhe Hu, Yinbo Chen, Jiarui Xu, Shubhankar Borse, Hong Cai, Fatih Porikli, and Xiaolong Wang. Learning implicit feature alignment function for semantic segmentation. In *ECCV*, 2022. 5

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[10] Hongjae Lee, Changwoo Han, and Seung-Won Jung. Gps-glass: Learning nighttime semantic segmentation using daytime video and gps data. *arXiv preprint arXiv:2207.13297*, 2022. 2

[11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. 1

[12] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 4

[13] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, 2020. 5

[14] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 1

[15] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 1, 2

[16] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *TPAMI*, 2020. 1, 2

[17] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, 2021. 1, 2

[18] Xin Tan, Ke Xu, Ying Cao, Yiheng Zhang, Lizhuang Ma, and Rynson WH Lau. Night-time scene parsing with a large real dataset. *TIP*, 30:9085–9098, 2021. 1, 2

[19] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *WACV*, 2021. 1, 2

[20] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018. 2

[21] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019. 2

[22] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *CVPR*, 2021. 2

[23] Xinyi Wu, Zhenyao Wu, Lili Ju, and Song Wang. A one-stage domain adaptation network with image alignment for unsupervised nighttime semantic segmentation. *TPAMI*, 2021. 2

[24] Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Ultrasr: Spatial encoding is a missing key for implicit image function-based arbitrary-scale super-resolution. *arXiv preprint arXiv:2103.12716*, 2021. 5

[25] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *CVPR*, 2021. 2

[26] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2

[27] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *IJCV*, 129(4):1106–1120, 2021. 1, 2

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