Unsupervised Modality Adaptation with Text-to-Image Diffusion Models for Semantic Segmentation

Ruihao Xia[1,2*], **Yu Liang**[2], **Peng-Tao Jiang**[2], **Hao Zhang**[2]
Bo Li[2†], **Yang Tang**[1,3†], **Pan Zhou**[4]
[1]East China University of Science and Technology, [2]vivo Mobile Communication Co., Ltd
[3]Peng Cheng Laboratory, [4]Singapore Management University

Abstract

Despite their success, unsupervised domain adaptation methods for semantic segmentation primarily focus on adaptation between image domains and do not utilize other abundant visual modalities like depth, infrared and event. This limitation hinders their performance and restricts their application in real-world multimodal scenarios. To address this issue, we propose Modality Adaptation with text-to-image Diffusion Models (MADM) for semantic segmentation task which utilizes text-to-image diffusion models pre-trained on extensive image-text pairs to enhance the model's cross-modality capabilities. Specifically, MADM comprises two key complementary components to tackle major challenges. First, due to the large modality gap, using one modal data to generate pseudo labels for another modality suffers from a significant drop in accuracy. To address this, MADM designs diffusion-based pseudo-label generation which adds latent noise to stabilize pseudo-labels and enhance label accuracy. Second, to overcome the limitations of latent low-resolution features in diffusion models, MADM introduces the label palette and latent regression which converts one-hot encoded labels into the RGB form by palette and regresses them in the latent space, thus ensuring the pre-trained decoder for up-sampling to obtain fine-grained features. Extensive experimental results demonstrate that MADM achieves state-of-the-art adaptation performance across various modality tasks, including images to *depth*, *infrared*, and *event* modalities. We open-source our code and models at https://github.com/XiaRho/MADM.

1 Introduction

Unsupervised Domain Adaptation for Semantic Segmentation (UDASS) involves a source domain with image-label pairs and a target domain with only unlabeled samples [9–11], and has achieved promising segmentation results in the image modality. Currently, most existing UDASS methods are restricted to transferring knowledge between similar image domains, such as from virtual scene [2, 12] to real scene [13], or from daytime scene [13] to nighttime scene [3, 4]. However, these approaches do not account for the wide range of visual modalities present in real-world scenarios, such as depth, infrared, and event modalities, which are valuable in nighttime perception [14–16] but often lack sufficient and high-quality labels for supervising segmentation training. Hence, in this paper, we are particularly interested in extending UDASS to Unsupervised Modality Adaptation for Semantic Segmentation (UMASS) across different visual modalities, i.e., the adaptation of a model from a labeled source image modality to an unlabeled target modality.

Differences across images arise from the objects, lighting, camera parameters, etc, while there are fundamental disparities in imaging principles across modalities that lead to greater variability. This

[*]Work was done during interning at vivo.
[†]Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

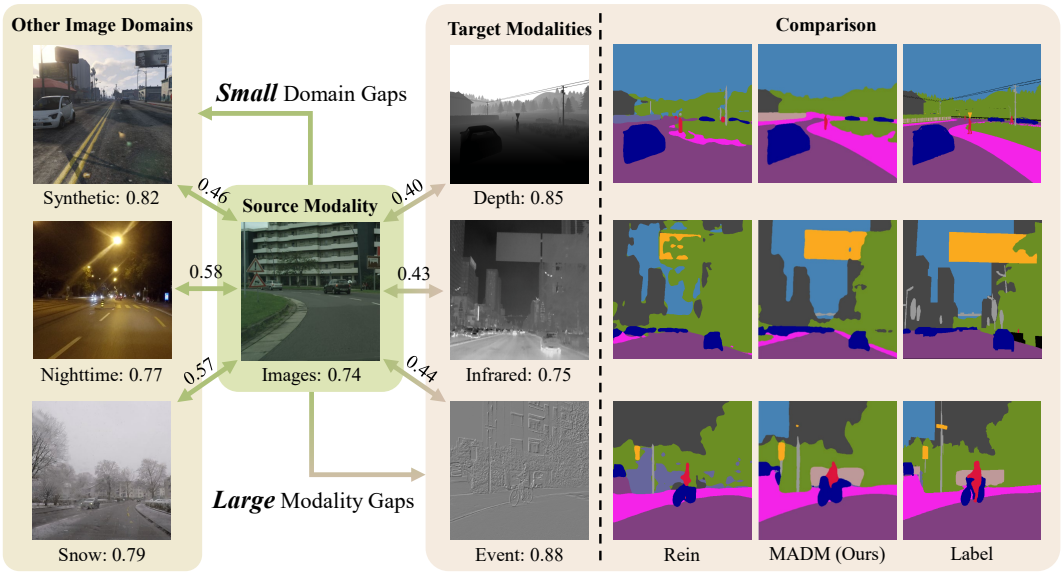

Figure 1: (1) On the left, we leverage the multi-modality model ImageBind [1] to quantify the similarity of images and modalities across datasets, i.e., GTA5-Synthetic [2], Dark Zurich-Nighttime [3], ACDC-Snow [4], DELIVER-Depth [5], FMB-Infrared [6], and DSEC-Event [7]. Specifically, we randomly select 500 samples from each dataset, and compute the average cosine similarity of the output vectors within the dataset (right side of the text) and between the datasets (on the arrows). (2) On the right, we compare the quantitative results with the state-of-the-art (SoTA) method Rein [8] on three different modalities.

is illustrated by Figure 1 which shows that the similarity among various image domains tends to be higher than between different modalities. These significant modality discrepancies poses significant challenges to existing UDASS methods [17, 18] on multimodal segmentation due to their limited pre-trained knowledge. Specifically, the backbone [19] used in current SoTA UDASS methods is pre-trained on the ImageNet-1K dataset [20] which contains one million images categorized into 1,000 distinct classes, providing the network with a foundational level of semantic understanding. While this backbone achieves promising results in UDASS, its insufficient pre-trained knowledge limits generalization to other visual modalities.

To address this issue, inspired by Text-to-Image Diffusion Models (TIDMs) [21], which are trained with internet-scale image-text pairs [22], we recognize that extensive pre-training data unifies samples with different distributions but similar semantic properties through texts, significantly enhancing the model's semantic understanding and generalization. Although TIDMs are not trained on other visual modalities, their large-scale samples and unification through texts enable them to adapt to a broader distribution of domains. Also, the extensive pretraining provides TIDMs with a robust understanding of high-level visual concepts, enabling their application to various domains, such as semantic matching [23], depth estimation [24], and 3D awareness [25]. This strong prior motivates us to utilize TIDMs as a robust backbone for solving UMA. Therefore, we present MADM: Modality Adaptation with text-to-image Diffusion Models, which takes full advantage of the generalization of pre-trained diffusion models and facilitates robust adaptation for accurate semantic segmentation in other visual modalities. Specifically, TIDM is used to extract robust features for segmentation and is trained in a self-training manner [17] which takes unlabeled target samples as input and generates pseudo-labels for training TIDM itself. Building upon TIDMs [21] and self-training [17], our MADM incorporates two innovative components: Diffusion-based Pseudo-Label Generation (DPLG) and Label Palette and Latent Regression (LPLR), which address the challenges of unstable pseudo-labeling and lack of fine-grained features extraction, respectively.

First, significant modality discrepancies hinder robust and high-quality pseudo-label generation, which is crucial for further training and enhancing the model. Directly using these unstable labels to train the model can lead to serious biases in the target modality. Thus, we propose DPLG which adds latent noise to target samples before generating pseudo-labels where the noise level gradually

decreases as training stabilizes. These pseudo-labels then supervise the noise-free target modality predictions. The mechanism behind DPLG leverages the denoising property of diffusion models, making the target modality more consistent with pre-trained inputs, thus improving accuracy. Unlike previous supervised diffusion-based semantic segmentation methods [26–28] which adopt a single-step forward and remove the diffusion process, we find that a proper diffusion process can stabilize the generation of pseudo-labels and adapt more successfully to the target modality.

Second, TIDMs [21] encode images into the latent space using a pre-trained Variational AutoEncoder (VAE) for diffusion and denoising, and then up-sample the latent output to the original resolution using the VAE decoder. When adopting TIDMs as a backbone, the resolution of the features is too low, resulting in a loss of details. To address this, we propose LPLR to convert pixel-level classification into regression in RGB form, utilizing the up-sampling capability of the pre-trained VAE decoder in a recycling manner. Specifically, we use a palette to convert one-hot encoded labels into RGB form and encode these RGB labels into latent representations. Then, the model is trained with a regression loss to fit the latent labels, obtaining high-resolution fine-grained features via the VAE decoder. Different from previous diffusion-based semantic segmentation methods [26–28] that directly extract multi-scale features from the denoising UNet network, we take inspiration from depth estimation with TIDMs [24] and propose LPLR to extract high-resolution features. Our contributions are summarized as follows:

- We propose MADM, extending traditional UDASS to UMASS with a pre-trained TIDM backbone for generalizing across various visual modalities.
- We design Diffusion-based Pseudo-Label Generation (DPLG) to provide more robust pseudo-labels by adding annealed latent noise to target samples for stable modality adaptation.
- We introduce Label Palette and Latent Regression (LPLR) to convert semantic segmentation into regression for learning details, thereby repeatedly utilizing pre-trained VAE decoders for high-resolution feature extraction.
- We demonstrate the effectiveness of our MADM through extensive experiments on three different modalities: Image [13] $\rightarrow$ Depth [5], Infrared [6], and Event [7].

2 Related Works

2.1 Unsupervised Domain Adaptation Semantic Segmentation

UDASS can be broadly categorized into two primary methods: adversarial and self-training methods. Adversarial approaches aim to align the distributions of the source and target domains at the level of images [29, 30] or features [31, 32] or outputs [33, 34], thereby facilitating the transfer of knowledge. On the other hand, self-training methods [9, 35] operate on the paradigm of pseudo-labeling, where the model's predictions on the target domain act as ground truth during training, iteratively refining the segmentation. Recently, the field has witnessed the development from CNN-based methods [30] to Transformer-based methods [17, 18, 36], leveraging the self-attention mechanism to capture long-range dependencies and enhance cross-domain feature representation.

However, most of these methods have predominantly focused on adaptation between different image domains, such as from synthetic [2, 12] to real-world images or across varying environmental conditions [4]. They will fail when adapting to other visual modalities, such as depth, infrared, or event, which have distinct data distribution. Thus, we propose MADM to address the limitation of UDASS for adapting to other unexplored visual modalities. Besides, different from multi/cross-modality domain adaptation [37, 38] which has paired two modalities in both domains, our modalities are different in source and target.

2.2 Text-to-Image Diffusion Models

Diffusion denoising probabilistic models [39] have set new benchmarks in the quality and controllability of generative tasks. These models are distinguished by their two-phase paradigm: the diffusion process that progressively adds noise to the sample, and the denoising process that learns to denoise the corrupted sample by a network. Utilizing the MSE loss between the residual noise and the prediction as a training objective, diffusion models have demonstrated greater training stability compared to generative adversarial networks [40] and VAEs [41]. To achieve high-quality controllable image

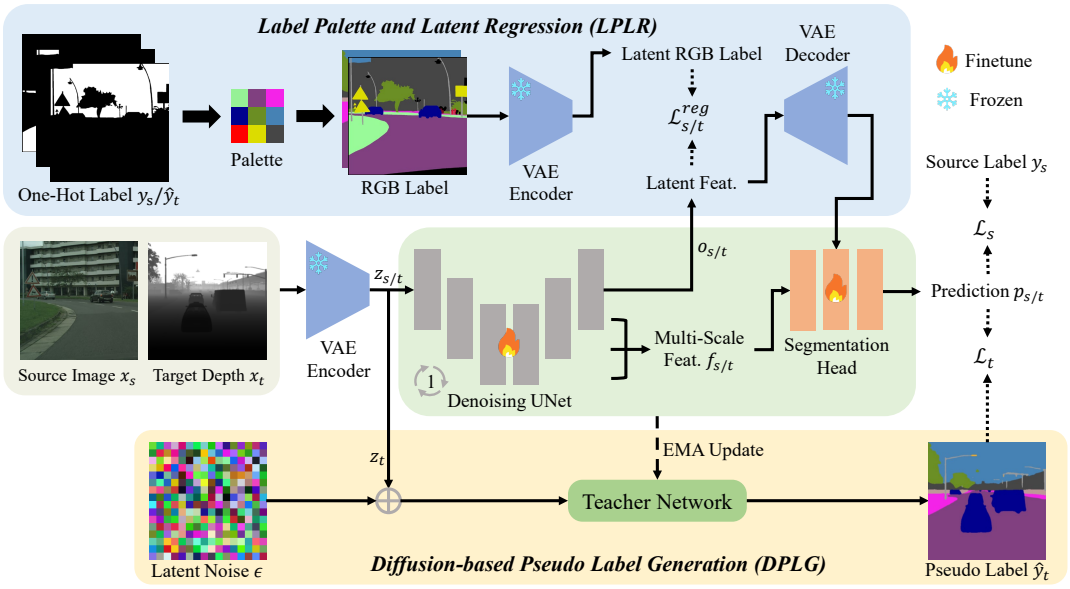

Figure 2: Our framework is divided into three parts. (1) Self-Training: Supervised loss in the source modality $\mathcal{L}_s$ and pseudo-labeled loss $\mathcal{L}_t$ in the target modality are used to train the network. (2) Diffusion-based Pseudo-Label Generation (DPLG): In the early stage of training, we add noise on the latent representation z_t to stabilize the pseudo-label generation. (3) Label Palette and Latent Regression (LPLR): The one-hot encoded labels $y_s/\hat{y}_t$ are converted to RGB form by palette and then encoded to the latent space to supervise the UNet output $o_{s/t}$.

generation with reduced computational demands, Rombach *et al.* [21] proposed the latent diffusion model that leverages cross-attention layers and confines the diffusion process to a low-resolution latent space, which has emerged as a widely recognized TIDM.

In recent advancements beyond the generative tasks, the exceptional semantic comprehension capabilities of TIDMs have been harnessed to enhance performance in many downstream applications. Xu *et al.* [27] presented a novel framework that integrates a pretrained TIDM with a discriminative model to address the challenges of open-vocabulary segmentation. Similarly, Zhao *et al.* [26] demonstrated the versatility of TIDMs by fine-tuning it to deal with various visual perception tasks, including semantic segmentation, referring image segmentation, and depth estimation. Gong *et al.* [28] introduced innovative scene prompts and a prompt randomization strategy on TIDMs, achieving new milestones in domain generalization and test-time domain adaptation. Their works highlight the potential of TIDMs to generalize across diverse domains and adapt to new ones with minimal additional training. It's worth noting that the aforementioned methods necessitate only a single-step forward pass through the denoising UNet, significantly streamlining the inference process.

The successful and diverse applications of TIDMs inspire our exploration into the generalization of TIDMs to more challenging visual modalities. However, we observe that the latent diffusion property within TIDMs leads to the lower-resolution feature extraction. To address this limitation, we propose the LPLR that converts semantic segmentation into latent regression to obtain fine-grained features.

3 Method

3.1 Overview

In UMASS, given the labeled source RGB modality $\{(x_s, y_s)\}$ and the unlabeled target modality $\{(x_t)\}$, our objective is to train a network which accepts x_t as input and outputs the corresponding semantic segmentation results p_t. As shown in Sec. 1, the primary challenge in this task stems from the significant disparities between the two modalities. To address this challenge, we propose to leverage the TIDM [21] as our backbone which is pre-trained on a vast array of image-text pairs [22] to enhance its generalization and can robustly extract features across modalities. Next, to overcome

the inaccurate pseudo-labels due to large modality gaps, we propose Diffusion-based Pseudo-Label Generation (DPLG), which stabilizes pseudo-label generation by injecting noise to the target modality. Moreover, TIDMs can only extract low-resolution features within the latent space, leading to a loss of semantic detail. To address this, we propose the Label Palette and Latent Regression (LPLR), which transforms pixel-wise classification into the regression, thereby allowing us to harness the fine-grained features upsampled by the pre-trained VAE Decoder. Our framework is illustrated in Figure 2. Next, we will introduce the our proposed DPLG and LPLR in detail.

3.2 Diffusion-based Pseudo-Label Generation

As shown in Figure 2, in MADM, we employ the TIDM to perform a single-step diffusion to extract multi-scale features from the intermediate output of the denoising UNet, following the approach in [26, 27]. First, samples from source and target modalities x_s, x_t are encoded to the latent representation $z_{s/t} = \mathcal{E}(x_{s/t})$ by the pretrained VAE encoder $\mathcal{E}$. Without any additional noise, we feed them to the denoising UNet and obtain multi-scale features $f_{s/t} = \mathrm{UNet}(z_{s/t}, c)$, where c is a learnable conditioned embedding instead of a textual description. Then, these features are subsequently fed into a segmentation head to generate the semantic segmentation prediction $p_{s/t} = \mathrm{Seg}(f_{s/t})$.

Our training method is anchored on the self-training DAFormer [17] prevalent in UDASS. The training objective is a composite of supervised loss from the source modality and pseudo-labeled loss from the target modality. For the labeled source modality, we use a cross-entropy loss between the prediction p_s and the ground truth labels y_s:

$$\mathcal{L}_s = \mathcal{L}_{CE}(p_s, y_s).$$

For the unlabeled target modality, we adopt a student-teacher paradigm in self-training [17]. Here, the existing network acts as the student, and through the Exponential Moving Average (EMA), we derive a teacher network [42]. The teacher network generates pseudo-labels $\hat{y}_t$, which then supervise the student network's output on $A(x_t)$, where $A(\cdot)$ represents the strong data augmentation [43].

However, the data distribution varies greatly between modalities. The teacher network is unable to provide accurate pseudo-labels for self-training, resulting in an unstable modality adaptation to the target modality. We observe that more robust pseudo-labels can be generated by injecting appropriate noise in the latent space and therefore propose the Diffusion-based Pseudo-Label Generation (DPLG). The proposed DPLG exploits the denoising property (perception of noise) of diffusion models to improve the robustness of semantic understanding on target samples.

Specifically, given a target sample x_t, we first encode it into the latent representation $z_t = \mathcal{E}(x_t)$ and then apply a diffusion process that adds noise ϵ to z_t to obtain a noisy latent representation:

$$z'_t = \sqrt{\bar{\alpha}_k} z_t + (1 - \sqrt{\bar{\alpha}_k})\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad k = \beta \cdot \max(0, 1 - i/\gamma).$$

Here, $\bar{\alpha}_k$ is a predetermined schedule that controls the amount of noise added at each step [39]. k is the diffusion step that controls the proportion of noise based on the initial diffusion step β and noise addition period γ, and i is the current iteration count.

This noisy representation z'_t is then used to generate pseudo-labels $\hat{y}_t = \mathrm{Seg}(\mathrm{UNet}(z'_t, c))$. In the pre-training of TIDMs [21], the objective is to estimate noise from latent inputs containing various noise levels. By injecting noise into the latent code, we effectively simulate this noisy distribution. This simulation aligns the latent space more closely with the data distribution encountered during the pre-training phase. Such alignment fosters a more robust and accurate semantic interpretation, which, in turn, enhances the quality of the pseudo labels generated. This shares similar

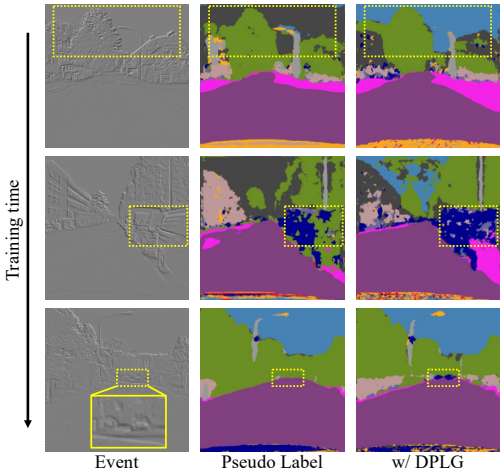

Figure 3: We visualize the pseudo-labels for event modality at the iteration of 1250, 1750, and 2250. The introduction of DPLG effectively improves the quality of pseudo-labels.

spirits in other applications of diffusion models, such as the text-to-3D [44] where injecting extra noise into data can improve the denoising quality of image and yields better pseudo labels. As shown in Figure 3, our DPLG can generate more accurate pseudo-labels compared to the noise-free addition. By strategically incorporating noise in the pseudo-label generation, DPLG enhances the model's adaptability to the target domain and mitigates the bias of semantic understanding. Then, the pseudo-labeled loss is formulated as:

$$\mathcal{L}_t = \mathcal{L}_{CE}(p_t, \hat{y}_t) \cdot q.$$

Here, q is the confidence value calculated by the softmax probability of p_t [17, 18, 43]. The consistency regularization of x_t between the teacher and student networks promotes adaptation to the target modality. It encourages the student network to match the teacher's predictions, even under the perturbations introduced by strong data augmentation.

3.3 Label Palette and Latent Regression

TIDMs compress the sample into a latent space with an 8x down-sampling factor, which leads to a serious loss of semantic detail. Specifically, for the input sample with a resolution of 512×512, it is reduced to 64×64 after embedded via $\mathcal{E}$. Then, within the denoising UNet decoder, multi-scale features are extracted $f_{s/t}$ after the 5th, 8th, and 11th blocks: 64×64, 32×32, and 16×16.

For diffusion-based depth estimation [24], leveraging the VAE decoder $\mathcal{D}$ to upsample the denoised latent representation back to the original resolution is a natural fit, which recovers the fine-grained scene details. However, the above method is not applicable to semantic segmentation due to the inherent difference between regression and classification. Therefore, to address the problem of semantic detail loss, we propose the Label Palette and Latent Regression (LPLR) module which converts semantic segmentation into regression and utilizes the VAE Decoder $\mathcal{D}$ to obtain high-resolution semantic features.

Initially, the one-hot encoded labels y_s and $\hat{y}_t$ are transformed into a perceptually meaningful RGB space with a pre-defined palette. These RGB representations are then encoded back into the latent space and supervise the UNet's output $o_{s/t} = \mathrm{UNet}(z_{s/t}, c)$ in a regression form:

$$\mathcal{L}^{reg}_{s/t} = |\mathcal{E}(\mathrm{Palette}(y_s/\hat{y}_t)) - o_{s/t}|.$$

With this supervision, we are able to utilize the VAE decoder $\mathcal{D}$ to obtain a high-resolution semantic regression feature $\mathcal{D}(o_{s/t})$. This feature, combined with the multi-scale features, is then fed to the segmentation head $p_{s/t} = \mathrm{Seg}(f_{s/t}, \mathcal{D}(o_{s/t}))$.

By employing LPLR, we effectively convert the semantic segmentation into a regression problem that can be tackled by the VAE decoder's upsampling capabilities, which retain the fine-grained details necessary for accurate segmentation. Finally, the total training objective is a sum of these losses:

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_t + \lambda_{reg}(\mathcal{L}^{reg}_s + \mathcal{L}^{reg}_t).$$

4 Experiments

4.1 Implementation Details

In our work, we utilize the Stable Diffusion v1-4 model [21], which has been pre-trained on the LAION-5B [22] dataset, as our TIDM. For the segmentation head, we instantiate it with the decoder from DAFormer [17]. We train our MADM for 10k iterations with a batch size of 2 and an image resolution of 512×512. The optimization is instantiated with AdamW [45] with a learning rate of 5e-6. For hyperparameters β, γ, and λ_{reg} in DPLG and LPLR, we set them to {5000,60,1.0}/{8000,50,1.0}/{8000,50,10.0} for depth/infrared/event modalities, respectively. In our experiments, we adopt the Cityscapes-Image [13] dataset as the source modality and the DELIVER-Depth [5], FMB-Infrared [6], and DSEC-Event [7] datasets as the target modalities. Since the semantic classes in these datasets are not identical, we merge some semantically similar classes during training. Experiments are conducted on a NVIDIA H800 GPU, occupying about 57G memory.

Table 1: Semantic segmentation results evaluated with MIoU (%) on three modalities. **Bold** numbers are the best, underscored second best.

(a) Cityscapes [13] → DELIVER-Depth [5].

Method	Sky	Build.	Fence	Person	Pole	Road	S.walk	Veg.	Vehi.	Wall	Tr.S.	MIoU (avg)
DAFormer [17]	82.28	43.35	11.82	56.03	13.90	80.10	15.44	60.08	72.67	0.18	**44.20**	43.64
MIC [18]	85.10	77.78	7.30	33.41	21.14	77.04	27.24	67.07	57.25	0.00	43.92	45.21
PiPa [46]	76.90	79.65	15.61	**60.21**	18.76	77.71	35.30	59.76	**84.54**	0.00	31.04	49.04
Rein [8]	92.00	78.78	**27.75**	43.88	**32.34**	78.81	27.50	58.06	76.45	**0.34**	36.68	50.23
MADM	**95.52**	**86.70**	12.48	41.88	18.99	**93.97**	**54.12**	**67.12**	84.29	0.00	33.34	**53.49**

(b) Cityscapes [13] → FMB-Infrared [6].

Method	Sky	Build.	Person	Pole	Road	S.walk	Veg.	Vehi.	Tr.S.	MIoU (avg)
DAFormer [17]	36.97	66.78	51.42	18.91	41.23	28.81	43.88	69.44	12.71	41.13
PiPa [46]	25.42	71.60	63.62	16.40	39.53	31.64	45.21	70.25	41.38	45.01
MIC [18]	38.11	71.63	57.89	17.59	40.68	**33.93**	49.49	70.26	29.85	45.49
Rein [8]	84.07	**72.84**	67.10	**26.40**	85.92	30.50	**72.61**	**84.51**	21.95	60.65
MADM	**88.79**	71.52	**70.51**	22.30	**89.08**	19.88	69.83	77.10	**51.08**	**62.23**

(c) Cityscapes [13] → DSEC-Event [7].

Method	Sky	Build.	Fence	Person	Pole	Road	S.walk	Veg.	Vehi.	Wall	Tr.S.	MIoU (avg)
DAFormer [17]	81.14	51.43	1.15	0.03	10.59	72.49	26.45	61.14	39.79	0.00	24.84	33.55
PiPa [46]	91.38	76.30	6.41	0.71	18.15	83.97	33.22	77.88	55.61	0.00	32.49	43.28
MIC [18]	92.36	**79.20**	6.69	**32.80**	19.30	79.75	31.46	68.17	58.35	0.01	**39.30**	46.13
Rein [8]	85.40	73.34	9.49	32.28	18.71	90.64	53.88	75.42	**79.44**	12.77	39.13	51.86
MADM	**92.60**	78.21	**26.51**	29.08	**22.78**	**92.20**	**62.90**	**81.70**	75.11	**23.92**	34.43	**56.31**

4.2 Datasets Setting

Cityscapes–Image. Cityscapes [13] is the source dataset in our experiments, which constitutes a real-world collection of street-view images captured across 50 distinct urban environments. The dataset is split into 2,975 training images and 500 validation images with a resolution of 2048×1024. It provides comprehensive semantic labeling at the pixel-level with 19 distinct semantic classes.

DELIVER–Depth. DELIVER [5] is a synthetic dataset containing five environmental conditions created by the CARLA simulator [47]. The dataset contains 25 semantic classes and 3,983/2,005/1,897 samples for training/validation/testing with a resolution of 1024×1024.

FMB–Infrared. FMB [6] is an urban street dataset with 1,500 RGB-Infrared pairs at a resolution of 800×600 with 14 semantic classes. It contains a wide range of real driving scenes under different lighting and weather conditions.

DSEC–Event. DSEC [7] is a stereo event camera dataset for driving scenarios. Driving data are recorded for 3,193 seconds in diverse illumination conditions and urban/rural environments. Event data have a resolution of 640×480 with 11 semantic classes and we aggregate them into the edge form in a recurrent manner [48].

4.3 Comparison with State of the Art Methods

Table 1 presents the comparison with existing SoTA methods DAFormer [17], MIC [18], PiPa [46], and Rein [8] across three modalities: Depth, Infrared, and Event. The comparison is based on the Mean Intersection over Union (MIoU) over all classes, a standard measure of segmentation accuracy.

Our MADM demonstrates a strong performance, achieving the MIoU of 53.49%, 62.23%, and 56.31% on the depth, infrared, and event modalities, respectively. It showcases a significant improvement over the SoTA method Rein [8] by +3.26%, +1.58%, +4.45%, which underscores the robustness and effectiveness of MADM in handling the other visual modalities. Also, it is worth noting that the self-training loss in our MADM is built upon DAFormer [17], exceeding it average of +17.9%.

Figure 4 offers a intuitive comparison of the semantic segmentation results. MIC [18] leverages the SegFormer backbone [19] that is pre-trained on the ImageNet-1k dataset [20], enabling it to capture

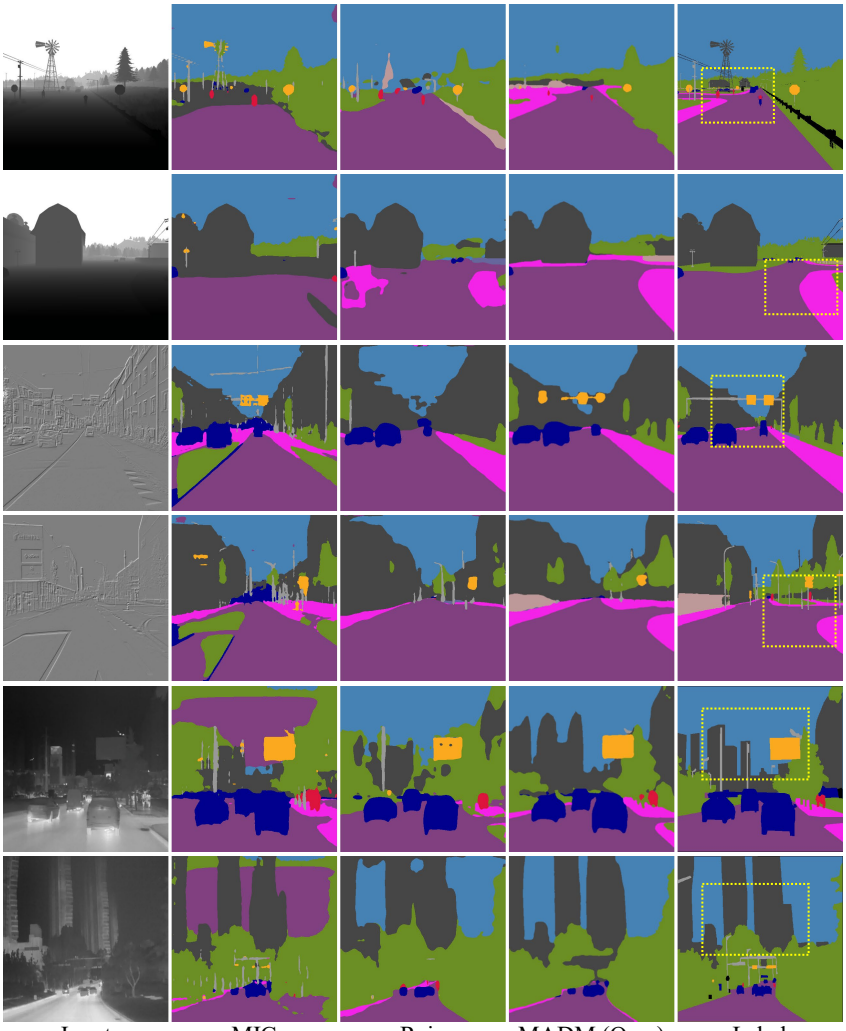

Figure 4: Qualitative semantic segmentation results generated by SoTA methods MIC [18], Rein [8], and our proposed MADM on three modalities.

more details within scenes, such as "pole". However, it exhibits weaker modality understanding, leading to frequent mis-segmentation, such as incorrectly classifying the "sky" as the "road". In contrast, Rein [8] is built upon the DINOv2 backbone that is pre-trained on extensive, curated datasets without explicit supervision [49]. This results in an improved semantic understanding of modalities compared to MIC [18]. Nonetheless, Rein still encounters issues with mis-segmentation and instances of under-segmentation.

Our MADM stands out for its exceptional ability to output precise segmentation results that closely mirror the ground truth. The incorporation of TIDM significantly enhances the generalization of our approach, providing an enhanced comprehension of diverse visual modalities and substantially mitigating the mis-segmentation.

4.4 Ablation Studies

Table 2 presents the complete ablation studies that quantify the performance gains achieved by incorporating our proposed DPLG and LPLR into the baseline. The "Baseline" column indicates the performance of the MADM model without DPLG and LPLR. It serves as a refer-

Table 2: Ablation of DPLG and LPLR in depth, infrared, and event modalities.

Modality	Baseline	w/ DPLG	w/ LPLR	MADM
Depth	50.61	51.65	52.91	**53.17**±0.26
Infrared	56.28	61.86	58.75	**62.14**±0.18
Event	52.27	52.84	53.05	**56.12**±0.20
Average	53.05	+2.40	+1.85	+4.09

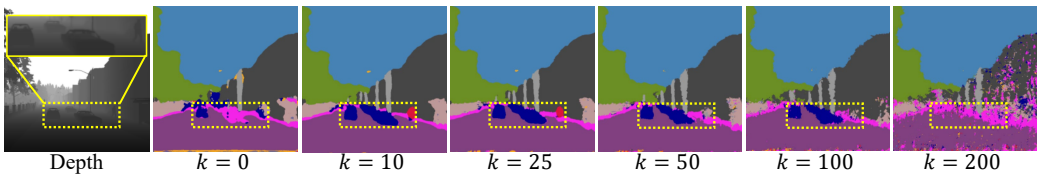

Figure 5: At the 1,250th iteration, we present a visual analysis of diffusion step k in DPLG.

ence point but achieves performance on par with the SoTA methods in Table 1, which demonstrates the strong generalization of TIDMs.

(1) When DGLP is employed, the MIoU is improved by an average of +2.40%, highlighting the effectiveness of generating robust pseudo-labels. Especially in the infrared modality, it achieves a +5.58% relative improvement over the baseline. The quantitative results of the pseudo-labels enhancement are shown in Figure 3. (2) The application of LPLR contributes to an average gain of +1.85%, emphasizing the importance of high-resolution features for segmentation tasks. (3) By employing both DGLP and LPLR, we observe a significant enhancement in +4.09% over the baseline, which underscores the synergistic benefits of combining robust pseudo-labels generation with fine-grained feature extraction.

4.5 Diffusion-based Pseudo-Label Generation

We analyze the pivotal roles of β and γ in our proposed DPLG, particularly within the depth modality. These two parameters control the diffusion step k on z_t, which is central to the stability and quality of pseudo-labels.

Table 3: Ablation of β and γ in DPLG on the depth modality.

β \ γ	2,000	5,000	8,000	Average
40	51.62	52.46	52.52	52.20
60	51.30	**53.49**	52.55	52.45
80	52.01	<u>52.85</u>	48.63	51.16
Average	51.64	52.93	51.23	-

Table 3 provides a detailed presentation of the impact of β and γ. For instance, when γ is set to 5,000, an increase in β from 40 to 60 leads to a noticeable improvement in performance, with the model achieving its peak score of 53.49%. However, further increasing β to 80 results in a decline in performance, indicating the existence of an optimal balance between these parameters.

In Figure 5, we offer an illustration of how the diffusion step k influences the generation of pseudo-labels. With noise-free addition ($k = 0$), the model encounters difficulties in accurately segmenting the "car" and "person" classes. Upon introducing a moderate quantity of noise ($k = 10 \sim 50$), the segmentation is noticeably enhanced, yielding more robust segmentation. Conversely, an excessive amount of noise ($k = 200$) leads to a significant degradation in segmentation.

4.6 Label Palette and Latent Regression

Table 4 analyzes the loss weight λ_{reg} within the proposed LPLR, which regulates the contribution of the regression losses. A minimal λ_{reg} of 1.0 and 3.0 yields MIoU of 53.31% and 54.40%, respectively, indicating the initial benefits of incorporating regression losses. Increasing λ_{reg} to 10.0 achieves the optimal MIoU of 56.31%, signifying the most effective balance between the segmentation and regression losses.

Table 4: Ablation of λ_{reg} in LPLR on event modality.

λ_{reg}	1.0	3.0	5.0	10.0	15.0
MIoU	53.31	54.40	<u>55.84</u>	**56.31**	55.31

In Figure 6, we offer an illustration of the impact of LPLR. It can be seen that the utilization of LPLR results in a more fine-grained segmentation, e.g., "person" and "vegetation" in the left yellow box, "road" and "sidewalk" in the right yellow box, which greatly improves the performance of our MADM.

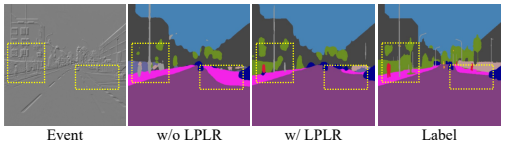

Figure 6: Qualitative Visualization on the event modality w/o and w/ our proposed LPLR. The prediction with LPLR shows more accurate fine-grained segmentation.

4.7 Benefits of MADM in Nighttime Datasets

As mentioned in Section 1, we indicate that other visual modalities present in real-world scenarios are valuable in nighttime perception. In this section, experiments are conducted on the infrared modality to prove this. The FMB-Infrared dataset [6] includes both image and infrared modalities on daytime and nighttime scenes. We adapt from cityscapes [13] with daytime RGB images to the nighttime image modality and infrared modality by our proposed MADA, respectively. Figure 7 and Table 5 show that the infrared modality has a clear advantage in the "Person" class due to obvious thermal differences and a good suppression of light interference.

Table 5: Semantic segmentation results of RGB and infrared modalities evaluated with MIoU (%) on FMB dataset [6].

Method	Sky	Build.	Person	Pole	Road	S.walk	Veg.	Vehi.	Tr.S.	MIoU (avg)
RGB	**88.85**	68.14	64.79	**25.80**	**89.09**	**32.43**	70.32	**84.13**	7.27	58.98
Infrared	87.94	**82.40**	**82.69**	21.50	76.21	26.50	**76.61**	83.80	**16.69**	**61.59**

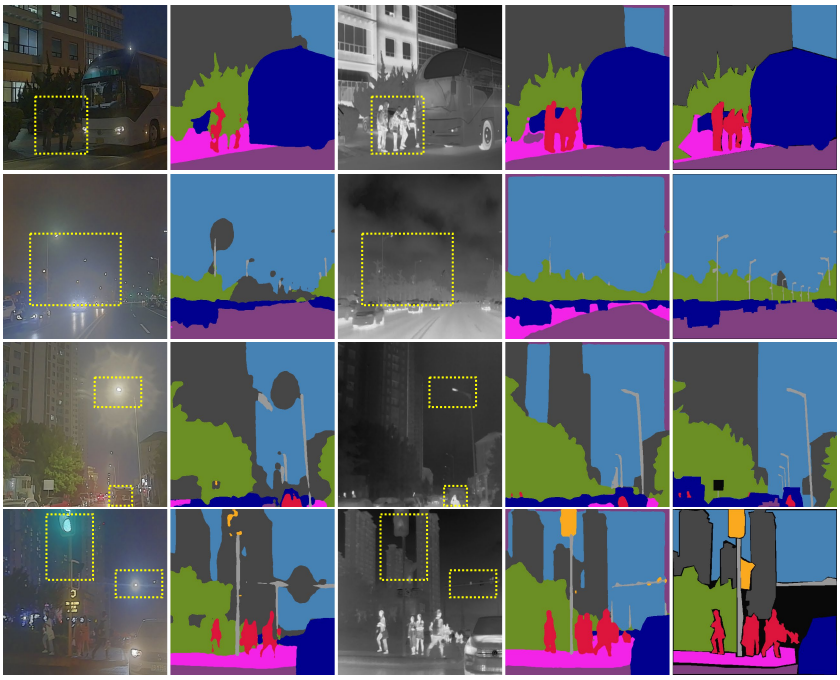

Figure 7: Visualization of daytime RGB images in Cityscapes dataset [13] → nighttime RGB and Infrared modalities in FMB dataset [6]

5 Conclusion

In this paper, we present MADM. With the powerful generalization of TIDMs, we extend domain adaptation to modality adaptation, aiming to segment other unexplored visual modalities in the real-world. Meanwhile, we propose DPLG and LPLR to solve the problems of pseudo-labeling instability and low-resolution features extraction within TIDMs. We hope our method can motivate further research on visual modalities other than RGB images. **Limitations:** However, despite using only a single-step forward for the diffusion model, the computation far exceeds existing UDASS networks. Future work could focus on distilling the knowledge of TIDMs into lightweight models when adapting. **Broader Impacts:** Our work pushes the boundary of semantic segmentation for other visual modalities, which will benefit several applications like multimodal fusion. To the best of our ability, MADM has little to no negative social impact.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Basic Science Center Program: 61988101), National Natural Science Foundation of China (62233005, 62293502), Fundamental Research Funds for the Central Universities(222202417006), the Programme of Introducing Talents of Discipline to Universities (the 111 Project) under Grant B17017. Pan Zhou was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grants (project ID: 23-SIS-SMU-028 and 23-SIS-SMU-070).

References

[1] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 2

[2] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision*, pages 102–118. Springer, 2016. 1, 2, 3

[3] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7374–7383, 2019. 1, 2

[4] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775, 2021. 1, 2, 3

[5] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1136–1147, 2023. 2, 3, 6, 7

[6] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8115–8124, 2023. 2, 3, 6, 7, 10

[7] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. DSEC: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. 2, 3, 6, 7, 14

[8] Zhixiang Wei, Lin Chen, Yi Jin, Xiaoxiao Ma, Tianle Liu, Pengyang Ling, Ben Wang, Huaian Chen, and Jinjin Zheng. Stronger, fewer, & superior: Harnessing vision foundation models for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 7, 8, 15

[9] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 3

[10] Quanliang Wu and Huajun Liu. Unsupervised domain adaptation for semantic segmentation using depth distribution. *Advances in Neural Information Processing Systems*, 35:14374–14387, 2022.

[11] Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, and Alexander Hauptmann. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. *Advances in Neural Information Processing Systems*, 33:3569–3580, 2020. 1

[12] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3234–3243, 2016. 1, 3

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 1, 3, 6, 7, 10

[14] Zhiwei Zhang, Zhizhong Zhang, Qian Yu, Ran Yi, Yuan Xie, and Lizhuang Ma. Lidar-camera panoptic segmentation via geometry-consistent and semantic-aware alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3662–3671, 2023. 1

[15] Johan Vertens, Jannik Zürn, and Wolfram Burgard. Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 8461–8468, 2020.

[16] Ruihao Xia, Chaoqiang Zhao, Meng Zheng, Ziyan Wu, Qiyu Sun, and Yang Tang. CMDA: Cross-modality domain adaptation for nighttime semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21572–21581, 2023. 1

[17] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9924–9935, 2022. 2, 3, 5, 6, 7, 14, 15

[18] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. MIC: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11721–11732, 2023. 2, 3, 6, 7, 8, 15

[19] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 2, 7

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015. 2, 7

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3, 4, 5, 6

[22] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2, 4, 6

[23] Xinghui Li, Jingyi Lu, Kai Han, and Victor Adrian Prisacariu. SD4Match: Learning to prompt stable diffusion model for semantic matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27558–27568, 2024. 2

[24] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3, 6

[25] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3D awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21795–21806, 2024. 2

[26] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5729–5739, 2023. 3, 4, 5

[27] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 4, 5

[28] Rui Gong, Martin Danelljan, Han Sun, Julio Delgado Mangas, and Luc Van Gool. Prompting diffusion representations for cross-domain semantic segmentation. *arXiv preprint arXiv:2307.02138*, 2023. 3, 4

[29] Suhyeon Lee, Junhyuk Hyun, Hongje Seong, and Euntai Kim. Unsupervised domain adaptation for semantic segmentation by content transfer. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 35, pages 8306–8315, 2021. 3

[30] Yunan Liu, Shanshan Zhang, Yang Li, and Jian Yang. Learning to adapt via latent domains for adaptive semantic segmentation. *Advances in Neural Information Processing Systems*, 34:1167–1178, 2021. 3

[31] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7892–7901, 2018. 3

[32] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 627–636, 2019. 3

[33] Matteo Biasetton, Umberto Michieli, Gianluca Agresti, and Pietro Zanuttigh. Unsupervised domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 3

[34] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEE Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018. 3

[35] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision*, pages 289–305, 2018. 3

[36] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. In *European Conference on Computer Vision*, pages 372–391. Springer, 2022. 3

[37] Duo Peng, Yinjie Lei, Wen Li, Pingping Zhang, and Yulan Guo. Sparse-to-dense feature matching: Intra and inter domain cross-modal learning in domain adaptation for 3D semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7108–7117, 2021. 3

[38] Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13618–13627, 2021. 3

[39] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3, 5

[40] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. 3

[41] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[42] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30, 2017. 5

[43] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. DACS: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1379–1389, 2021. 5, 6

[44] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 6

[45] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[46] Mu Chen, Zhedong Zheng, Yi Yang, and Tat-Seng Chua. Pipa: Pixel-and patch-wise self-supervised learning for domain adaptative semantic segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1905–1914, 2023. 7, 15

[47] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Conference on Robot Learning*, pages 1–16. PMLR, 2017. 7

[48] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3857–3866, 2019. 7

[49] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 8

A Appendix

A.1 Visualization of LPLR

We visualize LPLR under different iteration steps in Figure 8. "Regression" and "Classification" in Figure 8 denote the output of the VAE decoder and segmentation head, respectively. Our proposed LPLR leverages the up-sampling capability of a pre-trained VAE decoder in a recycling manner. As the model converges, the regression results transform from blurry to progressively clearer states, presenting more details compared to the classification results. This assists the segmentation head in producing more accurate semantic segmentation results.

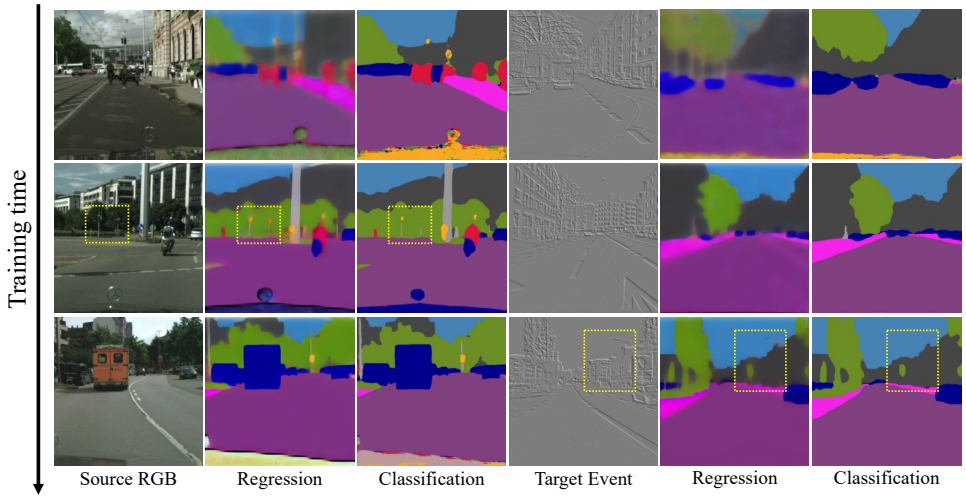

Figure 8: Visualization of the output of VAE decoder (Regression) and segmentation head (Classification).

A.2 Influence of Different Data Volumes

Table 6: Influence on different data volumes testing on the DSEC dataset [7].

Method	Baseline-100%	MADM-10%	MADM-25%	MADM-50%	MADM-100%
MIoU	52.27	53.21	53.69	54.55	**56.31**

We train our method with 10%, 25%, and 50% of the total target samples in the event modality. Here, the "Baseline-100%" column indicates the performance of the MADM model without DPLG and LPLR and trained on the whole target samples. The results in Table indicate that our proposed MADM consistently outperforms the baseline across all tested data volumes. Additionally, our MADM is robust and effective even when the dataset size is relatively small.

A.3 Parameters and Costs

Table 7 presents a detailed comparison of training timeper iteration, number of iterations, total training time, parameters, and performance across various methods in the DSEC event modality [7], including our MADM and its distilled variant.

While MADM does exhibit a higher training time per iteration, the advanced visual prior derived from TIDMs necessitates fewer iteration for adaptation, presenting a minimum total training time. Moreover, MADM achieves a substantial performance improvement, with an MIoU of 57.34%, surpassing other methods. Recognizing the trade-off in parameter count, we have leveraged our MADM model as a teacher to perform a secondary self-training. This approach has enabled us to distill the knowledge embedded in MADM into a more compact DAFormer model [17], MADM

Table 7: Comparison of parameters and costs.

Method	Training/Iter. (seconds)	Iteration	Total training (hours)	Params (million)	MIoU
DAFormer [17]	0.36	40k	4.0	85	33.55
PiPa [46]	1.12	60k	18.7	85	43.28
MIC [18]	0.48	40k	5.3	85	46.13
Rein [8]	1.25	40k	13.9	328	51.86
MADM	1.38	10k	3.8	949	**56.31**
MADM (Distilled)	0.46	10k	1.3	85	54.03

(Distilled), which retains a highMIoU of 54.03% while significantly reducing parameters to 85M and only increasing the training time by 1.3 hours. Our distilled model demonstrates that it is possible to maintain high performance with reduced computational costs, addressing the concerns raised regarding the parameters and efficiency of MADM.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction of the paper provide a clear and concise overview of the research's main contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in the conclusion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our work does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We detail our work in the Methods section and describe implementation details in the Experiments section.

   Guidelines: The paper thoroughly details all the necessary components for reproducing the main experimental results.

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will release the code as soon as our work is accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We specify all the training and test details in the implementation details section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report error bars with overall run with given experimental conditions.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the computer resources in the implementation details section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our research conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impacts our work in the conclusion.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly credite the creator and cite the original paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not introduce new assets in this paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.