RAFT: ***R***obust ***A***ugmentation of ***F***ea***T***ures for Image Segmentation

Anonymous CVPR submission

Paper ID *****

Abstract

Image segmentation is a powerful computer vision technique for scene understanding. However, real-world deployment is stymied by the need for high-quality, meticulously labeled datasets. Synthetic data provides high-quality labels while reducing the need for manual data collection and annotation. However, deep neural networks trained on synthetic data often face the Syn2Real problem, leading to poor performance in real-world deployments.

*To mitigate the aforementioned gap in image segmentation, we propose **RAFT**, a novel framework for adapting image segmentation models using minimal labeled real-world data through data and feature augmentations, as well as active learning. To validate RAFT, we perform experiments on the synthetic-to-real "SYNTHIA→Cityscapes" and "GTAV→Cityscapes" benchmarks. We manage to surpass the previous state of the art, HALO. SYNTHIA→Cityscapes experiences an improvement in mIoU* upon domain adaptation of 2.1%/79.9%, and GTAV→Cityscapes experiences a 0.4%/78.2% improvement in mIoU. Furthermore, we test our approach on the real-to-real benchmark of "Cityscapes→ACDC", and again surpass HALO, with a gain in mIoU upon adaptation of 1.3%/73.2%. Finally, we examine the effect of the allocated annotation budget and various components of RAFT upon the final transfer mIoU.*

1. Introduction

Image segmentation is a fundamental task in computer vision and digital image processing, which attempts to separate an image into discrete regions on a per-class basis. It involves classifying pixels or groups of pixels based on shared characteristics such as intensity, color, texture, or spatial proximity, serving as a critical preprocessing step in applications such as medical imaging[25] , autonomous driving[17], remote sensing[11], and robot navigation[1, 40]. Deep neural networks have revolutionized image segmentation, becoming the state-of-the-art approach for this task [26]. Despite these advancements, their reliance on large datasets with pixel-level annotations is a major barrier to widespread deployment. Automatically generated synthetic data presents a promising solution, enabling the creation of virtually unlimited datasets covering diverse scenarios at minimal cost. However, models trained on synthetic data often struggle to generalize to real-world data—an issue commonly known as the Syn2Real or Sim2Real problem[16].

Specifically, the Sim2Real and Syn2Real problems are a subset of the larger domain shift problem [29]. Although well-curated synthetic datasets may share the same semantic content as a real-world counterpart, they often do not share the same "style" [32]. Within the task of computer vision, given the difficulty of achieving complete photorealism and accuracy within images, synthetic data generally contains simplified geometry, textures, and lighting compared to its real-world counterparts. Therefore, when training a model mostly or exclusively on synthetic data, the distribution of image features the model learns from ends up differing significantly from the distribution of image features within real-world domains.

Most techniques proposed to tackle this domain shift attempt to reduce the distance between the training distribution and the target distribution. For example, the current state of the art in domain adaptation for image segmentation is Hyperbolic Active Learning Optimization [8] (HALO), which takes advantage of properties of hyperbolic geometry to perform active domain adaptation [43] [30]. Through carefully curated label acquisition of a small percentage of especially challenging pixels from the real-world domain, HALO creates a hybrid training distribution closer to the target distribution. However, while this strategy offers significant benefits for underrepresented classes, it also inherently limits how much the training distribution can be shifted.

To this end, we propose: **R**obust **A**ugmentation of **F**ea**T**ures for Image Segmentation, or **RAFT**. RAFT extends HALO's method of active domain adaptation for expanding the synthetic training distribution with a minimal amount of real-world data and feature augmentation. We showcase a high-level overview of RAFT in Figure 1. Hyperbolic feature augmentation steadily expands the distribution of each class by generating novel features within those

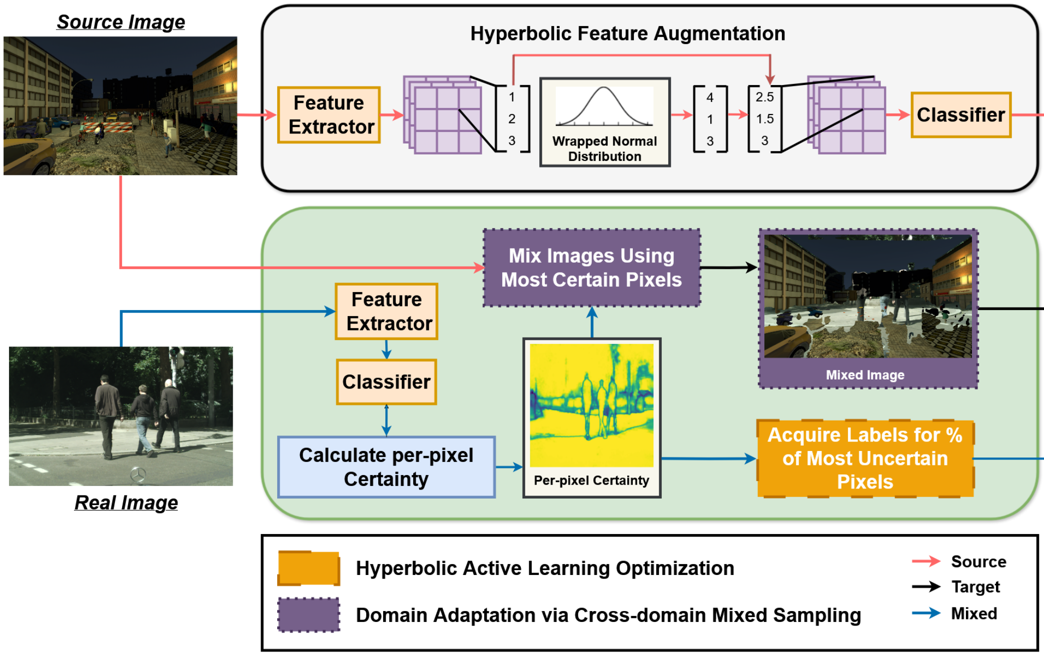
Source Image
Hyperbolic Feature Augmentation
Feature Extractor
1
2
3
Wrapped Normal Distribution
4
1
3
2.5
1.5
3
Classifier
Mix Images Using Most Certain Pixels
Feature Extractor
Classifier
Calculate per-pixel Certainty
Per-pixel Certainty
Mixed Image
RAFT Loss
Acquire Labels for % of Most Uncertain Pixels
Real Image
Hyperbolic Active Learning Optimization
Domain Adaptation via Cross-domain Mixed Sampling
Source
Target
Mixed

Figure 1. The proposed architecture of our RAFT framework. The classifier allows for active learning via uncertainty detection, while the HFA module generates novel instances of classes, thus enabling better generalization upon domain transfer.

classes, through sampling and interpolation. Our implementation of Domain Adaptation via Cross-domain Mixed Sampling [36] (DACS) utilizes the same combined hyperbolic radius and entropy certainty measure HALO does in order to select pixels of the target dataset in which the model has a high degree of prediction certainty. From these regions of high certainty, pseudolabels are generated, and the source-domain image has its pixels replaced with those of the pixels of high certainty from the target dataset. Thus, a combined source-target image along with corresponding labels is generated during training time. Our contributions are summarized as follows:

- We extend Hyperbolic Feature Augmentation (HFA) from image classification to image segmentation tasks.
- We utilize the uncertainty predictions HALO gives us to perform Domain Adaptation via Cross-Domain Mixed Sampling (DACS).
- We achieve state of the art results on the SYNTHIA→Cityscapes and GTAV→Cityscapes benchmarks. On SYNTHIA→Cityscapes, we achieve a 2.1% and 1.4% improvement in the 13-class and 16-class mIoU's respectively upon adaptation, leading to a 79.9% 13-class and 83.5% 16-class mIoU, respectively upon transfer. We get a smaller but still notable improvement on GTAV→Cityscapes of 0.4%, leading to a final mIoU of 78.2%
- We evaluate RAFT on the real-to-real domain adaptation benchmark of Cityscapes→ACDC and achieve an improvement in mIoU upon adaptation of 1.3%, resulting in a final mIoU of 73.2%
- We ablate each component of our proposed RAFT framework and their contribution to the final model performance upon domain adaptation.

2. Related Works

In this section, we will address the key challenges relevant to our RAFT framework by reviewing related work on area imbalance, augmentation methods, and knowledge distillation strategies in semantic segmentation.

2.1. Handling Data Imbalance

Data imbalance in image segmentation occurs when dominant regions, such as large backgrounds, overwhelm smaller, critical objects, leading to suboptimal model training and poor performance on rare classes.

2.1.1. Algorithmic Approaches

To address this, algorithm-based approaches like the localized maximum likelihood decision rules by Chan et al. [5] reweight pixel predictions to better detect rare classes, Remote sensing imagery, with its multi-scale and complex scenes, presents additional challenges. Recent works have integrated scale-adaptive mechanisms within network architectures to tackle these issues. For example, Wang et al. [39] designed an unbalanced class learning network that dynamically fuses multi-scale features, and Zhou et al. [45] introduced a dynamic effective class balanced approach using weighting strategies based on effective samples.

2.1.2. Data Augmentation Methods

In addition to algorithmic approaches, data augmentation is another technique for dealing with data imbalance. Traditional augmentation methods, such as geometric transformations (rotation, flipping, cropping, and scaling) and photometric adjustments (brightness, contrast, and color alterations), have been widely used to improve model generalization by increasing dataset diversity [2]. Deep learning-based augmentation methods leverage generative models [14, 15, 33] to synthesize new data points that capture complex variations beyond simple transformations. These generative models have been successfully applied in medical imaging [4, 9, 24, 41] and underwater object recognition[21, 22], demonstrating their potential to generate realistic synthetic samples that enhance model robustness.

Recent advancements in augmentation strategies extend beyond raw image transformations to feature space modifications. Methods such as feature-based augmentation [20] [38] introduce diversity at a more abstract level, leveraging learned feature embeddings to generate novel training samples. This approach has been particularly effective in semi-supervised learning settings, where labeled data is scarce [19].

Hyperbolic neural networks often struggle to generalize when trained on few-shot, limited datasets. HFA addresses this issue by leveraging feature augmentations in hyperbolic space. Specifically, HFA generates class-identity-preserving features by modeling their distribution with a per-class wrapped normal distribution on the hyperbolic manifold. To accurately estimate the parameters of each distribution—including hyperbolic curvature, mean, and covariance—HFA employs a meta-learning framework based on neural ordinary differential equations (ODEs). In this framework, the iterative update of distribution parameters is modeled as a continuous gradient flow, which is then solved via the RK4 [3] method. This neural ODE-based gradient flow network leverages prior knowledge to achieve a more precise approximation of the underlying distribution even in data-scarce regimes.

Furthermore, a Euclidean upper bound on the augmentation loss is derived, negating the need for computationally expensive hyperbolic operations, and enabling efficient training of a distance-based classifier in hyperbolic space. These augmentation techniques, whether in image or feature space, collectively improve model generalization and robustness in real-world applications. Building on this, our RAFT framework extends Hyperbolic Feature Augmentation (HFA) from classification to segmentation tasks and integrates multiple complementary augmentations to explicitly address class imbalance and uncertainty.

3. Method

In this section, we introduce RAFT (Robust Augmentation of FeaTures), our framework for domain adaptation in image segmentation. We first provide an overview of HALO, which forms the foundation of our approach, and then present our novel extensions: (1) a pixel-level adaptation of Hyperbolic Feature Augmentation (HFA), (2) a hyperbolic mixup technique, (3) a class-balanced focal loss, and (4) Domain Adaptation via Cross-Domain Mixed Sampling (DACS). Together, these components form a comprehensive solution to the Syn2Real problem in image segmentation.

3.1. Hyperbolic Active Learning Optimization

HALO provides the foundation for our approach by leveraging hyperbolic geometry to identify data-scarce regions. It interprets hyperbolic radius—the distance of hyperbolic pixel embeddings from the origin of the hyperbolic space—as a proxy for data scarcity. By combining this radius with prediction entropy, HALO generates an acquisition score that guides active learning, identifying the most uncertain pixels for label acquisition.

The key insight of HALO is that by strategically acquiring labels for a small set of challenging real-world pixels and combining them with fully labeled synthetic data, the training distribution can be expanded to more closely match the target distribution. This reduces the domain gap in a label-efficient manner.

However, both HALO and other works have noted that rare or underrepresented classes (e.g., pedestrians and cyclists in autonomous driving datasets) exhibit disproportionately high classification uncertainty due to dataset class imbalance. This area imbalance problem limits the effectiveness of uncertainty-based active learning alone.

Our RAFT framework addresses this limitation by integrating HFA and DACS into HALO's active learning stage. These additions generate more diverse, challenging training data specifically for classes disadvantaged by area imbalance. As these augmented samples shift the training distribution over time and reduce its distance from the target distribution, the overall classification uncertainty decreases, allowing the acquisition stages of HALO to focus exclusively on the most challenging areas for label acquisition.

3.2. Pixel-Level Hyperbolic Feature Augmentation

A key challenge in adapting HFA from image classification to semantic segmentation lies in the fundamental differences between their feature spaces. In image classification, input images are heavily compressed into relatively simple feature vectors prior to classification, with each embedding ultimately representing a single class. This allows the simple neural ODE architectures in the original HFA to effectively model this restricted embedding space on a per-class basis.

In contrast, the dense feature maps created by image segmentation networks are significantly larger and more complex, retaining substantial spatial information and potentially containing data for many classes simultaneously. Generating such detailed feature maps while preserving accurate spatial information is beyond the capabilities of the original neural ODE approach.

Therefore, we take a different approach. Similar to the original HFA, we generate an approximate hyperbolic wrapped normal distribution for each semantic class via neural ODEs. However, instead of attempting to generate entire feature maps, we sample individual pixel embeddings from these class-specific distributions. Using these sampled pixel embeddings, we then perform weighted interpolation in hyperbolic space between the pixel embeddings extracted from the image and our generated embeddings on a per-class basis. To avoid confusion, throughout the rest of this subsection, we use the term ”real” as a shorthand for the pixel embeddings we extract from training images.

Specifically, we utilize the weighted Möbius gyromidpoint [37]:

$$m_\kappa(x_1, \dots, x_n, \alpha_1, \dots, \alpha_n) = \frac{1}{2} \otimes_\kappa \left(\sum_{i=1}^{n} \frac{\alpha_i \lambda_{x_i}^{\kappa}}{\sum_{j=1}^{n} \alpha_j \left(\lambda_{x_j}^{\kappa} - 1 \right)} x_i \right) \tag{1}$$

where x_i represents the real and sampled pixel embeddings, α_i represents the weight that each x_i contributes to the final interpolated embedding, and $-\kappa$ is the curvature of the hyperbolic space. Following HALO, we fix the curvature at -1.

To balance diversity and stability during training, we dynamically adjust the interpolation weights:

$$\alpha_i = \begin{cases} \alpha_{\text{initial}} - t \cdot \frac{\alpha_{\text{initial}} - \alpha_{\text{final}}}{T} & \text{if } x_i \text{ is real} \\ 1 - \left(\alpha_{\text{initial}} - t \cdot \frac{\alpha_{\text{initial}} - \alpha_{\text{final}}}{T} \right) & \text{if } x_i \text{ is sampled} \end{cases} \tag{2}$$

where t is the current training step, T is the total number of training steps, $\alpha_{\text{initial}} = 0.8$ is the initial weight for real embeddings, and $\alpha_{\text{final}} = 0.5$ is the final weight. This means we initially rely more heavily on real embeddings (80% real, 20% sampled), and gradually transition to more heavily weight the sampled embeddings over the course of training.

Unlike the original HFA, which uses a distance-based classifier and a Euclidean upper bound for its loss function, we retain HALO's hyperbolic multinomial logistic regression [10] (HyperMLR) pixel classifier.

3.2.1. Hyperbolic Mixup

To further increase feature diversity while preserving manifold structure, we implement mixup in hyperbolic space. For each class, we take real pixel embeddings $\{h_i\}_{i=1}^{n_j}$, and create pairs by shuffling them to obtain $\{h_i'\}_{i=1}^{n_j}$. We then sample coefficients $\lambda_i \sim \text{Beta}(\alpha, \alpha)$. Finally, we perform geodesic interpolation using the Möbius gyromidpoint:

$$\tilde{h}_i^{mix} = m_\kappa(h_i, h_i', \lambda_i, 1 - \lambda_i) \tag{3}$$

We combine these mixed embeddings with the sampled embeddings from our learned class distributions into a single augmentation pool $\tilde{H}_{aug} = [\tilde{h}^{mix}, \tilde{h}^{synth}]$. When reintegrating these features into the spatial feature map, we randomly select either the mixed or sampled embeddings on a per-class basis.

3.2.2. Class-Balanced Focal Loss

To directly address class imbalance in image segmentation, we integrate a class-balanced focal loss [27] adapted for hyperbolic space:

$$\mathcal{L}_{\text{CB}}(y, \hat{y}) = -\sum_{c=1}^{C} \frac{1 - \beta}{1 - \beta^{n_c}} (1 - p_c)^{\gamma} y_c \log(p_c) \tag{4}$$

where n_c is the number of pixels belonging to class c, β is a hyperparameter controlling class balancing, $\gamma = 2.0$ is the focusing parameter, and p_c is the predicted probability for class c. This approach automatically adjusts the weight of each class based on its frequency while focusing on hard-to-classify pixels, which is particularly beneficial for boundary regions and minority classes.

3.2.3. Meta-Learning for Distribution Estimation

Following the original HFA methodology, we use a meta-learning approach to train the gradient flow networks for distribution estimation. For each training iteration, we randomly partition the source dataset into a training set $\mathcal{D}_t$ and a validation set $\mathcal{D}_v$. Within the inner loop, we then use $\mathcal{D}_t$ to estimate distribution parameters via neural ODEs and train the segmentation model with generated augmentations. Finally, within the outer loop we evaluate model performance on $\mathcal{D}_v$ and update the gradient flow networks to minimize validation loss.

The complete HFA loss is formulated as:

$$\begin{aligned} \mathcal{L}_{\text{hfa}} = {} & \underbrace{\mathcal{L}_{\text{orig\_cls}} + \mathcal{L}_{\text{aug\_cls}}}_{\text{Classification Losses}} + \underbrace{\lambda_{\text{div}} \mathcal{L}_{\text{div}}}_{\text{Diversity Loss}} \\ & + \underbrace{\lambda_{\text{proto\_reg}} \mathcal{L}_{\text{proto\_reg}}}_{\text{Prototype Regularization}} + \underbrace{\lambda_{\text{mean\_var}} \mathcal{L}_{\text{mean\_var}}}_{\text{Distribution Regularization}} \end{aligned} \tag{5}$$

Where $\mathcal{L}_{\text{orig\_cls}}$ is the classification loss on original features, $\mathcal{L}_{\text{aug\_cls}}$ is the classification loss on augmented features, $\mathcal{L}_{\text{div}}$ promotes diversity, $\mathcal{L}_{\text{proto\_reg}}$ regularizes class prototype locations in hyperbolic space, and $\mathcal{L}_{\text{mean\_var}}$ constrains the estimated distribution parameters to prevent overfitting.

3.3. Domain Adaptation via Cross-Domain Mixed Sampling

The final component of our RAFT framework is Domain Adaptation via Cross-Domain Mixed Sampling (DACS). DACS enhances unsupervised domain adaptation by mixing labeled images from the source domain with unlabeled images from the target domain through a class-wise cut-and-paste approach.

The major innovation in our implementation is leveraging the certainty measures already computed by HALO. Specifically, we identify regions in the target dataset where the model has high prediction certainty (using the same combined hyperbolic radius and entropy measure that HALO uses). We then generate pseudo-labels for these high-certainty regions and replace corresponding pixels in the source domain image with these high-certainty pixels from the target domain. This creates a mixed source-target image with corresponding labels during training, effectively leveraging the most reliable information from the target domain.

3.4. Training Process

Following HALO's approach, we first pretrain our image segmentation model on the source dataset and then perform domain adaptation. During the domain adaptation stage, we retain HALO's mixed active learning/supervised learning approach but additionally apply our feature augmentations.

The final composite loss during active domain adaptation is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{src}} + \mathcal{L}_{\text{tgt}} + \lambda_{\text{hfa}}\mathcal{L}_{\text{hfa}} + \mathcal{L}_{\text{dacs}} \tag{6}$$

where λ_{hfa} is the weight assigned to the HFA loss, which we keep set to 0.1.

Through this comprehensive approach, RAFT effectively addresses both the domain gap and the class imbalance issues inherent in Syn2Real image segmentation, leading to significant performance improvements as demonstrated in our experiments.

4. Experimental Setup

4.1. Datasets

To evaluate our proposed RAFT framework, we first conduct experiments on the widely used synthetic-to-real domain adaptation benchmarks of **SYNTHIA→Cityscapes GTAV→Cityscapes**, as well as the real-to-real domain adaptation benchmark of **Cityscapes→ACDC**. SYNTHIA [34] and GTAV [31] contain 9,000 and 25,000 synthetic images respectively. On the other hand, Cityscapes [7] consists of 25,000 images captured from cars in various cities around Germany, with 5,000 of these images having fine-grained labels. Finally, ACDC [35] contains 4000 fine-grained labeled images captured from cars in adverse settings containing rain, snow, fog, and nighttime conditions.

4.2. Implementation Details

Within all of our experiments, we made use of PyTorch [28] to develop and train our models. To perform calculations in hyperbolic space, we made use of the geoopt [18] library, and in order to train the neural ODE's used for estimating the wrapped normal distribution parameters, we use the torchdiffeq [6] library. We resize all images from GTAV and SYNTHIA to 1280×720, while we resize all images from both Cityscapes and ACDC to 1280×640.

Due to its excellent performance in image segmentation tasks, we make use of the SegFormer [44] architecture. For our benchmarks, we specifically utilize the B4 variant of SegFormer variant, with 64.1 million parameters. For our ablation studies examining annotation budgets and RAFT components, we make use of the SegFormer B0 variant which contains only 3.7 million parameters. When training our SegFormer models, we utilize the AdamW [23] optimizer to train all components of our model and the HFA components, with a base learning rate of 6×10^{-5} and a polynomial schedule using a power of 0.5. For training the HFA components, we again use AdamW, however, we use a base learning rate of 6×10^{-6}, and no scheduler.

We evaluate our models via the standard metrics for image segmentation of mean Intersection-over-Union and per-class Intersection-over-Union. Apart from SYNTHIA→Cityscapes, each benchmark has 19 classes and we only report a singular mIoU value for these 19 classes. SYNTHIA has 16 classes and we report two mIoU metrics when evaluating a SYNTHIA-trained model, one for only 13 classes (mIoU), and one for all 16 classes (mIoU*).

5. Results

In this section, we describe the outcomes of our various benchmarks, and analyze the impact of various components within our proposed method.

5.1. Comparison With the State-of-the-Art

Table 1 shows the results of SYNTHIA→Cityscapes domain adaptation. RAFT's performance exceeds that of the other state-of-the-art methods, with a 13-class mIoU of 79.9, and a 16-class mIoU of 83.5%. Even using the same architecture and annotation budget, RAFT manages to improve over the previous best method, HALO, with an improvement in the 13-class mIoU of 2.1%, and an improvement in the 16-class mIoU of 1.4%. On GTAV→Cityscapes, with the same annotation budget, RAFT similarly displays an improvement, achieving a modest gain of 0.4% over HALO, resulting in a final mIoU of

Model	mIoU	mIoU*	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	sky	person	rider	car	bus	motorcycle	bicycle
RAFT SegFormer B4 (ours)	**79.9**	**83.5**	98.3	87.1	93.0	66.1	64.6	62.2	69.2	77.8	93.3	95.2	81.8	62.9	95.4	89.2	65.9	76.8
HALO SegFormer B4 [8]	77.8	82.1	98.3	86.5	92.6	61.0	61.5	60.6	67.6	76.2	93.2	94.6	80.8	58.9	95.0	85.1	62.7	75.6
RIPU DeepLabv2 [42]	70.1	75.7	96.8	76.6	89.6	45.0	47.7	45.0	53.0	62.5	90.6	92.7	73.0	52.9	93.1	80.5	52.4	70.1
ILM-ASSL DeepLabv3+ [12]	76.6	82.1	97.4	80.1	91.8	38.6	55.2	64.1	70.9	78.7	91.6	94.5	82.7	60.1	94.4	81.7	66.8	77.2
DWBA-ADA DeepLabv3+ [13]	72.7	78.1	97.4	90.3	47.2	47.9	53.4	57.2	67.6	91.7	94.2	76.2	55.0	93.8	83.4	55.1	72.1	78.1

Table 1. Comparison of Syn2Real methods for image segmentation on SYNTHIA to Cityscapes. mIoU* utilizes 13 classes, excluding ”wall”, ”fence”, and ”pole”, while mIoU utilizes all 16 classes within SYNTHIA.

Model	mIoU	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
RAFT SegFormer B4 (ours)	**78.2**	98.3	85.8	92.7	63.8	62.7	61.6	69.2	77.3	92.5	64.0	94.9	80.9	62.3	95.1	86.5	86.1	73.4	63.3	75.6
HALO SegFormer B4 [8]	77.8	98.2	85.4	92.5	62.5	61.6	58.3	67.7	74.9	92.2	65.1	94.7	79.9	60.8	94.6	84.1	85.4	83.6	61.2	75.5
RIPU DeepLabv2 [42]	71.2	97.0	77.3	90.4	54.6	53.2	47.7	55.9	64.1	90.2	59.2	93.2	75.0	54.8	92.7	73.0	79.7	68.9	55.5	70.3
ILM-ASSL DeepLabv3+ [12]	76.1	96.9	77.8	91.6	46.7	56.0	63.2	70.8	77.4	91.9	54.9	94.5	82.3	61.2	94.9	79.3	88.1	75.3	65.8	77.6
DWBA-ADA DeepLabv3+ [13]	71.9	97.5	80.5	90.8	54.7	52.2	53.3	55.7	65.2	91.0	61.0	93.5	75.3	53.6	92.9	81.8	75.2	62.9	57.8	71.6

Table 2. Comparison of Syn2Real methods for image segmentation on GTAV to Cityscapes

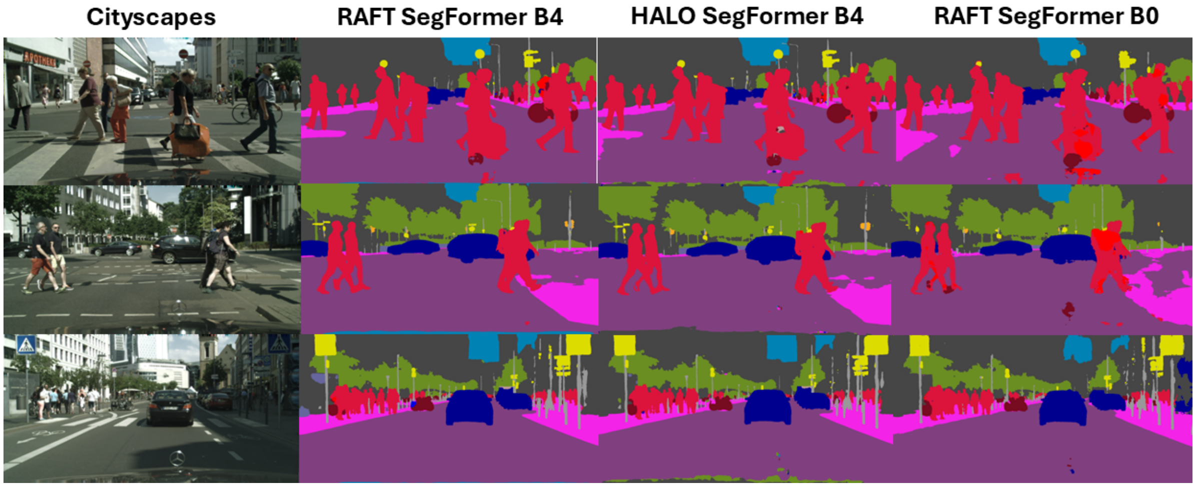

Figure 2. On the left are various images from Cityscapes’ validation split. In the middle are segmentation masks created by our SegFormer B4 model trained via our proposed RAFT framework. On the right are segmentation masks created by a DeepLabv3+ model trained via ILM-ASSL.

78.2% upon domain adaptation as shown in Table 2. Table 3 showcases the real-to-real Cityscapes→ACDC benchmark results, RAFT improves over HALO by 1.3%, with an mIoU upon transfer of 73.2%.

Examining the segmentation masks generated by RAFT and HALO using SegFormer B4 on the Cityscapes validation split, shown in Figure 2, while both segmentations are generally high quality, one notices that where HALO appears to struggle with the hood and its ornament of the car the photos are being captured from, RAFT has comparatively little trouble in ignoring it, with the obvious exception of the first photo, in which it misclassifies the hood ornament as being a bicycle, and a small part of the hood as being a person. Additionally, within the second photo,

Model	mIoU	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
RAFT SegFormer B4 (ours)	**73.2**	95.7	81.0	88.5	62.6	53.7	65.0	77.7	67.0	87.9	54.7	95.7	66.1	35.3	89.1	82.7	89.7	90.4	48.5	58.8
HALO SegFormer B4 [8]	71.9	95.2	79.8	88.2	60.2	51.1	64.1	78.2	65.6	87.9	55.7	95.5	66.3	20.7	88.9	82.2	89.3	87.9	50.4	59.0
RIPU DeepLabv3+ [42] [8]	63.5	92.7	72.5	84.7	53.1	44.8	56.7	69.1	58.9	85.9	46.9	95.3	57.2	24.3	84.5	61.4	59.4	79.0	36.9	43.6

Table 3. Comparison of active domain adaptation methods for image segmentation on Cityscapes to ACDC

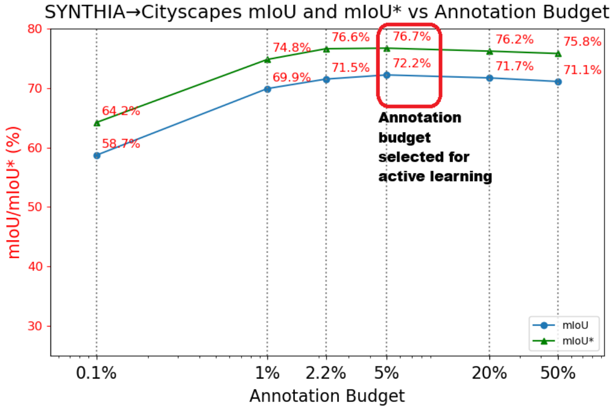

Figure 3. The effect on mIoU and mIoU* of allocating varying percentages of target domain labels for active domain adaptation from **SYNTHIA→Cityscapes**. The mIoU* metric uses 13 common classes in both SYNTHIA and Cityscapes, while the mIoU metric uses all 16 classes shared between SYNTHIA and Cityscapes. We found an annotation budget of 5% performed the best, with it achieving both the highest mIoU and mIoU* upon domain adaptation.

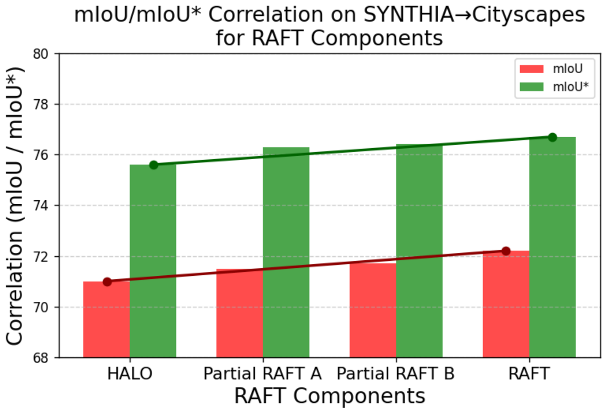

Figure 4. The effect on mIoU and mIoU* of applying various RAFT components in performing domain adaptation of a SegFormer B0 model from SYNTHIA to Cityscapes. The mIoU* metric uses 13 common classes in both SYNTHIA and Cityscapes, while the mIoU metric uses all 16 classes shared between SYNTHIA and Cityscapes. Partial RAFT A includes HALO along with HFA and hyperbolic mixup, Partial RAFT B includes the aforementioned components plus the class-balanced focal loss, and RAFT includes all RAFT components.

RAFT misclassifies a small sliver of the hood as being sky pixels. A smaller but still noticeable area where RAFT improves over HALO is in the sidewalk the pedestrians are walking to in photo 2. While the RAFT-trained SegFormer B4 still doesn't fullly classify the sidewalk correctly, it classifies more of the overall shape compared to the HALO-trained SegFormer B4.

These results confirm the effectiveness of RAFT in improving domain adaptation performance, without the need for additional labeled target domain data over HALO.

5.2. Annotation Budget

The annotation budget in active domain adaptation for image segmentation defines the total amount of labeling resources allocated for annotating target domain data. In the context of uncertainty-based active learning, this budget constrains the selection of the most uncertain pixels or regions for annotation, typically by specifying the proportion of high-uncertainty pixels to be labeled. The ideal outcome is that the amount of manual labeling effort is minimized, while maximizing model performance on the target domain. We experimented with a variety of different annotation budgets as shown in Figure 3, and similarly to HALO, found that 5% of the target domain labels gave us our best results when validating on Cityscapes. As a result, we fixed our annotation budget at 5% for all our other experiments.

5.3. RAFT Component Ablation

As our RAFT framework is composed of multiple components, we performed an ablation study evaluating the effect each component had on the final mIoU upon domain transfer using SYNTHIA → Cityscapes as our benchmark. As shown in Figure 4, each component played a role in the final RAFT mIoU upon domain transfer. Given that we build upon HALO, we use it as our baseline. With HALO alone, we achieve an mIoU and mIoU* of 71 and 75.6 respectively. We then combined HALO with our image segmentation-adapted HFA and hyperbolic mixup, which we call Partial RAFT A. This combination results in an mIoU and mIoU* of 71.5 and 76.3 respectively, or a 0.5% and 0.9% improvement over HALO alone. We further extend Partial RAFT A with the class-balanced focal loss, which we then call Partial RAFT B. This Partial RAFT B results in an mIoU and mIoU* of 71.7 and 76.4 respectively, or a modest 0.2% and 0.1% improvement over Partial Raft A. Finally, integrating this with DACS, giving us the full RAFT framework, results in an mIoU and mIoU* of 72.2 and 76.7 respectively, which is a 0.5% and 0.3% improvement over Partial RAFT B.

5.4. Per-Pixel Classification Uncertainty

Figure 5 showcases per-pixel uncertainty measures captured from RAFT and HALO-trained SegFormer B4 models on the same input image from the Cityscapes validation split. The lighter colors showcase areas of higher uncertainty, and vice versa. The RAFT-trained SegFormer B4 model show-

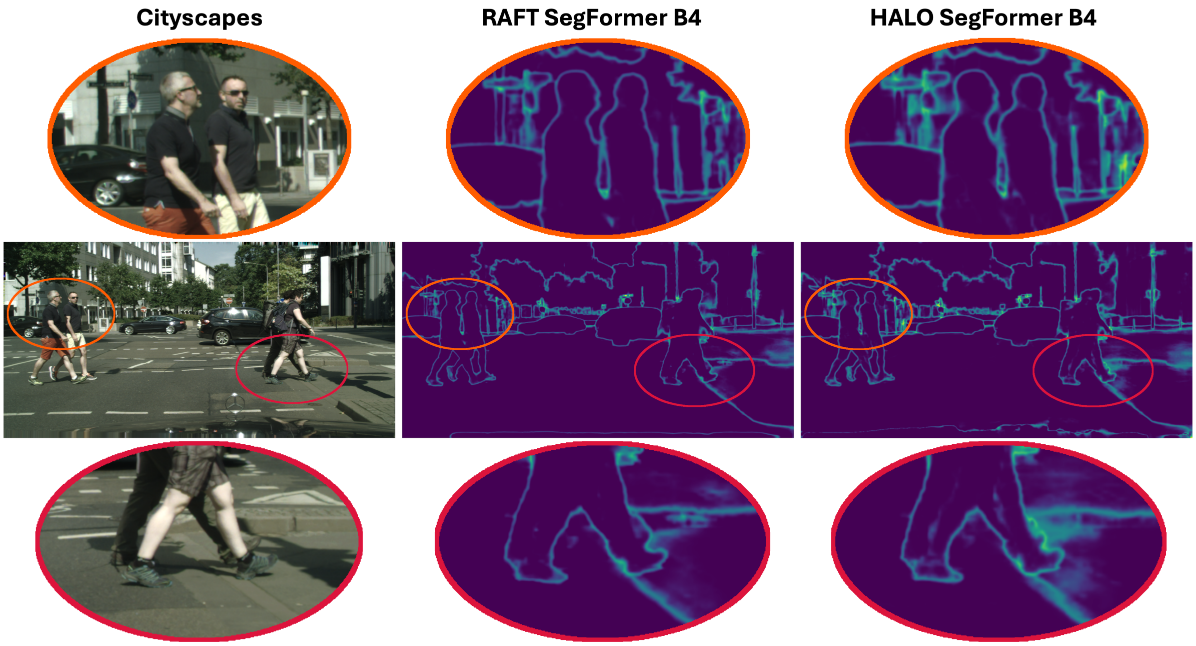

Figure 5. On the left are various images from Cityscapes' validation split, along with zoomed in . In the middle and on the right are measures of pixel classification uncertainty. The lighter the color, the higher the degree of uncertainty.

cases noticeably less classification uncertainty compared to its HALO counterpart. We zoom in on two areas with especially noticeable differences in uncertainty. Both areas showcase plenty of "fuzz" under the HALO-trained models - noticeable around the street signs by the two people walking from the torso up, as well as on the sidewalk by the two people walking from the legs down. Additionally, even highly uncertain areas show a lower degree of uncertainty compared to HALO, with lighter coloration and less fuzz.

Classifications having low uncertainty doesn't necessarily mean that the generated segmentation masks will be completely accurate. However, the lower uncertainty around classes negatively affected by area imbalance, such as street signs and people, does seem to indicate that RAFT achieved one of our intended effects in creating additional, diverse samples for these disfavored classes.

6. Conclusion

In this work, we introduced RAFT (**R**obust **A**ugmentation of **F**ea**T**ures), a novel framework that effectively addresses the Syn2Real problem in image segmentation through a combination of augmentation techniques and active learning. We verify that our framework effectively performs Syn2Real domain adaptation through experimentation on the SYNTHIA→Cityscapes and GTAV→Cityscapes benchmarks. RAFT achieves state-of-the-art performance, with notable improvements of 2.1% in 13-class mIoU (79.9%) and 1.4% in 16-class mIoU (83.5%) on SYNTHIA→Cityscapes, as well as a 0.4% improvement (78.2% mIoU) on GTAV→Cityscapes. Furthermore, RAFT's effectiveness extends to real-to-real domain adaptation, shown by our results on the Cityscapes→ACDC benchmark, where we achieve a 1.3% improvement (73.2% mIoU) over previous methods. Additionally, our ablation studies confirm that each component contributes meaningfully to the final performance, with the complete RAFT framework delivering superior results compared to partial implementations.

While RAFT advances the state-of-the-art in domain-adaptive image segmentation, several promising directions remain for future work. First, exploring the application of our techniques to other segmentation architectures beyond SegFormer could validate the general applicability of our approach. Additionally, in this work, we exclusively used synthetic data for creating the wrapped normal distributions we sample from in HFA, in future works, we could explore utilizing the small amount of labeled target data to generate the per-class wrapped normal distributions.

By advancing domain adaptive semantic segmentation through RAFT, we take an important step toward enabling more robust computer vision systems that can generalize effectively from synthetic training data to real-world environments, addressing a critical challenge in applying data-hungry image segmentation neural networks in the real-world.

References

[1] Adnan Abdullah, Titon Barua, Reagan Tibbetts, Zijie Chen, Md Jahidul Islam, and Ioannis Rekleitis. Caveseg: Deep semantic segmentation and scene parsing for autonomous underwater cave exploration. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3781–3788. IEEE, 2024. 1

[2] Khaled Alomar, Halil Ibrahim Aysel, and Xiaohao Cai. Data augmentation in classification and segmentation: A survey and new strategies. *Journal of Imaging*, 9(2):46, 2023. 3

[3] John Charles Butcher. A history of runge-kutta methods. *Applied numerical mathematics*, 20(3):247–260, 1996. 3

[4] Jinzheng Cai, Zizhao Zhang, Lei Cui, Yefeng Zheng, and Lin Yang. Towards cross-modal organ translation and segmentation: A cycle-and shape-consistent generative adversarial network. *Medical image analysis*, 52:174–184, 2019. 3

[5] Robin Chan, Matthias Rottmann, Fabian Hüger, Peter Schlicht, and Hanno Gottschalk. Application of decision rules for handling class imbalance in semantic segmentation. *arXiv preprint arXiv:1901.08394*, 2019. 2

[6] Ricky T. Q. Chen. torchdiffeq, 2018. 5

[7] 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. 5

[8] Luca Franco, Paolo Mandica, Konstantinos Kallidromitis, Devin Guillory, Yu-Teng Li, Trevor Darrell, and Fabio Galasso. Hyperbolic active learning for semantic segmentation under domain shift. *arXiv preprint arXiv:2306.11180*, 2023. 1, 6

[9] Michael Gadermayr, Laxmi Gupta, Vitus Appel, Peter Boor, Barbara M Klinkhammer, and Dorit Merhof. Generative adversarial networks for facilitating stain-independent supervised and unsupervised segmentation: a study on kidney histology. *IEEE transactions on medical imaging*, 38(10): 2293–2302, 2019. 3

[10] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *Advances in neural information processing systems*, 31, 2018. 4

[11] Akshaj Gaur, Cheng Liu, Xiaomin Lin, Nare Karapetyan, and Yiannis Aloimonos. Whale detection enhancement through synthetic satellite images. In *OCEANS 2023-MTS/IEEE US Gulf Coast*, pages 1–7. IEEE, 2023. 1

[12] Licong Guan and Xue Yuan. Iterative loop method combining active and semi-supervised learning for domain adaptive semantic segmentation. *arXiv preprint arXiv:2301.13361*, 2023. 6

[13] Licong Guan and Xue Yuan. Dynamic weighting and boundary-aware active domain adaptation for semantic segmentation in autonomous driving environment. *IEEE Transactions on Intelligent Transportation Systems*, 25(11): 18461–18471, 2024. 6

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018. 3

[16] Xuemin Hu, Shen Li, Tingyu Huang, Bo Tang, Rouxing Huai, and Long Chen. How simulation helps autonomous driving: A survey of sim2real, digital twins, and parallel intelligence. *IEEE Transactions on Intelligent Vehicles*, 9(1): 593–612, 2023. 1

[17] Çağrı Kaymak and Ayşegül Uçar. A brief survey and an application of semantic image segmentation for autonomous driving. *Handbook of deep learning applications*, pages 161–200, 2019. 1

[18] Max Kochurov, Rasul Karimov, and Serge Kozlukov. Geoopt: Riemannian optimization in pytorch, 2020. 5

[19] Chia-Wen Kuo, Chih-Yao Ma, Jia-Bin Huang, and Zsolt Kira. Featmatch: Feature-based augmentation for semi-supervised learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 479–495. Springer, 2020. 3

[20] Pan Li, Da Li, Wei Li, Shaogang Gong, Yanwei Fu, and Timothy M Hospedales. A simple feature augmentation for domain generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8886–8895, 2021. 3

[21] Xiaomin Lin, Nitin J Sanket, Nare Karapetyan, and Yiannis Aloimonos. Oysternet: Enhanced oyster detection using simulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5170–5176. IEEE, 2023. 3

[22] Xiaomin Lin, Vivek Mange, Arjun Suresh, Bernhard Neuberger, Aadi Palnitkar, Brendan Campbell, Alan Williams, Kleio Baxevani, Jeremy Mallette, Alhim Vera, et al. Odyssee: Oyster detection yielded by sensor systems on edge electronics. *arXiv preprint arXiv:2409.07003*, 2024. 3

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[24] Ali Madani, Mehdi Moradi, Alexandros Karargyris, and Tanveer Syeda-Mahmood. Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation. In *2018 IEEE 15th International symposium on biomedical imaging (ISBI 2018)*, pages 1038–1042. IEEE, 2018. 3

[25] Michael Maynord, M Mehdi Farhangi, Cornelia Fermüller, Yiannis Aloimonos, Gary Levine, Nicholas Petrick, Berkman Sahiner, and Aria Pezeshk. Semi-supervised training using cooperative labeling of weakly annotated data for nodule detection in chest ct. *Medical Physics*, 50(7):4255–4268, 2023. 1

[26] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021. 1

[27] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep

neural networks using focal loss. *Advances in neural information processing systems*, 33:15288–15299, 2020. 4

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[29] Xingchao Peng, Ben Usman, Neela Kaushik, Dequan Wang, Judy Hoffman, and Kate Saenko. Visda: A synthetic-to-real benchmark for visual domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2021–2026, 2018. 1

[30] Piyush Rai, Avishek Saha, Hal Daumé III, and Suresh Venkatasubramanian. Domain adaptation meets active learning. In *Proceedings of the NAACL HLT 2010 workshop on active learning for natural language processing*, pages 27–32, 2010. 1

[31] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision (ECCV)*, pages 102–118. Springer International Publishing, 2016. 5

[32] Pavel Rojtberg, Thomas Pöllabauer, and Arjan Kuijper. Style-transfer gans for bridging the domain gap in synthetic pose estimator training. In *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 188–195. IEEE, 2020. 1

[33] 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. 3

[34] 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. 5

[35] Christos Sakaridis, Haoran Wang, Ke Li, René Zurbrügg, Arpit Jadon, Wim Abbeloos, Daniel Olmeda Reino, Luc Van Gool, and Dengxin Dai. Acdc: The adverse conditions dataset with correspondences for robust semantic driving scene perception. *arXiv e-prints*, pages arXiv–2104, 2021. 5

[36] 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. 2

[37] Abraham Albert Ungar. *A gyrovector space approach to hyperbolic geometry*. Springer International Publishing, Cham, 2009. 4

[38] Riccardo Volpi, Pietro Morerio, Silvio Savarese, and Vittorio Murino. Adversarial feature augmentation for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5495–5504, 2018. 3

[39] He Wang, Mengmeng Zhang, Wei Li, Yunhao Gao, Yuanyuan Gui, and Yuxiang Zhang. Unbalanced class learning network with scale-adaptive perception for complicated scene in remote sensing images segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2

[40] Kasun Weerakoon, Adarsh Jagan Sathyamoorthy, Utsav Patel, and Dinesh Manocha. Terp: Reliable planning in uneven outdoor environments using deep reinforcement learning. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 9447–9453. IEEE, 2022. 1

[41] Thomas Wollmann, CS Eijkman, and Karl Rohr. Adversarial domain adaptation to improve automatic breast cancer grading in lymph nodes. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 582–585. IEEE, 2018. 3

[42] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, and Xinjing Cheng. Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8068–8078, 2022. 6

[43] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, Xinjing Cheng, and Guoren Wang. Active learning for domain adaptation: An energy-based approach. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8708–8716, 2022. 1

[44] 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. 5

[45] Zheng Zhou, Change Zheng, Xiaodong Liu, Ye Tian, Xiaoyi Chen, Xuexue Chen, and Zixun Dong. A dynamic effective class balanced approach for remote sensing imagery semantic segmentation of imbalanced data. *Remote Sensing*, 15(7):1768, 2023. 2