Learning Part Segmentation from Synthetic Animals

Jiawei Peng[1] Ju He[1] Prakhar Kaushik[1] Zihao Xiao[1]
Jiteng Mu[2] Alan Yuille[1]

[1]Johns Hopkins University [2]UC San Diego

Abstract

*Semantic part segmentation provides an intricate and interpretable understanding of an object, thereby benefiting numerous downstream tasks. However, the need for exhaustive annotations impedes its usage across diverse object types. This paper focuses on learning part segmentation from synthetic animals, leveraging the Skinned Multi-Animal Linear (SMAL) models to scale up existing synthetic data generated by computer-aided design (CAD) animal models. Compared to CAD models, SMAL models generate data with a wider range of poses observed in real-world scenarios. As a result, our first contribution is to construct a synthetic animal dataset of tigers and horses with more pose diversity, termed Synthetic Animal Parts (SAP). We then benchmark Syn-to-Real animal part segmentation from SAP to PartImageNet, namely SynRealPart, with existing semantic segmentation domain adaptation methods and further improve them as our second contribution. Concretely, we examine three Syn-to-Real adaptation methods but observe relative performance drop due to the innate difference between the two tasks. To address this, we propose a simple yet effective method called **C**lass-**B**alanced **F**ourier **D**ata **M**ixing (CB-FDM). Fourier Data Mixing aligns the spectral amplitudes of synthetic images with real images, thereby making the mixed images have more similar frequency content to real images. We further use Class-Balanced Pseudo-Label Re-Weighting to alleviate the imbalanced class distribution. We demonstrate the efficacy of CB-FDM on SynRealPart over previous methods with significant performance improvements. Remarkably, our third contribution is to reveal that the learned parts from synthetic tiger and horse are transferable across all quadrupeds in PartImageNet, further underscoring the utility and potential applications of animal part segmentation.*

1. Introduction

Semantic parts of an object provide a hierarchical representation which enables detailed and interpretable understanding of the object, which can facilitate various downstream tasks. For instance, humans can estimate the pose of a tiger based on the spatial configuration of its part and hence classify whether it is about to attack or lying down to rest. These hierarchical representations have also been proved to be important in many computer vision tasks, e.g., pose estimation [1, 2], detection [3, 4], segmentation [5, 6], fine-grained recognition [7]. However, the annotation of part segmentation on real images is very expensive, especially for general non-rigid objects, like animals. To the best of our knowledge, the only two datasets that offer animal part segmentation annotation are PASCAL-Part [8] and PartImageNet [9]. While these datasets offer accurate and valuable annotations, they are limited in number of animal samples and time-consuming to scale up to more species.

By contrast, annotating parts on synthetic data is a much cheaper way to achieve the goal of scalability. Prior research [10, 11] annotated parts on 3D computer-aided design (CAD) models and rendered synthetic images based on the CAD models. With automatic-generated ground truth, this methodology offers numerous advantages, primarily in significantly reducing annotation costs. Once annotated, it can generate arbitrary number of synthetic images from arbitrary viewpoints. However, this approach comes across challenges in animal part segmentation due to the pose diversity in these CAD models is limited and does not encompass the diverse poses observed in the natural world.

As illustrated in Fig 1, in this paper, we propose to expand the pose space for CAD data by fitting the Skinned Multi-Animal Linear (SMAL) models [12] with more poses and utilizing them to generate supplementary synthetic data. Similar with SMPL models [13], SMAL models build a parametric way to represent the animal shape and pose based on strong prior and is widely used in 3D animal pose and shape estimation. This process requires additional keypoints annotation and silhouette masks to reconstruct SMAL models from images. Inspired by [14], we replace the manual labeling process for silhouettes with the prediction of pre-trained object segmentation model [15]. Combining the new SMAL data with the previous CAD data, we construct a synthetic animal dataset with diverse pose configurations of tiger and

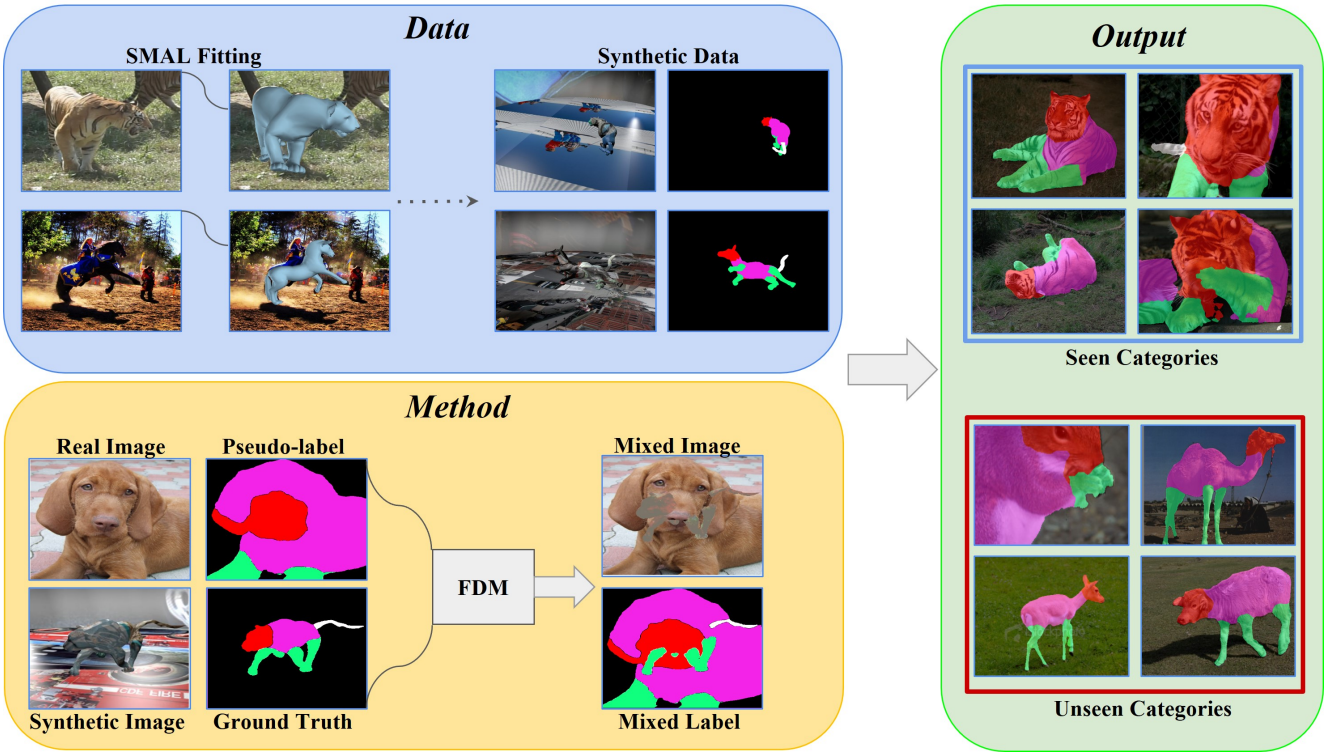

Figure 1. **Overview.** We generate synthetic animals by fitting SMAL models and rendering with random viewpoints and textures (top-left). We exploit Fourier Data Mixing (FDM) to align the spectral amplitudes of unlabeled real animals from PartImageNet and generated synthetic animals to obtain mixed images (bottom-left). Along with the Class-Balanced (CB) training strategy, our model is capable of segmenting real animals on seen categories (top-right). Moreover, the model is capable to transfer part knowledge to unseen categories (bottom-right).

horse, termed Synthetic Animal Parts (SAP). Then we set up a new Syn-to-Real benchmark of animal part segmentation called SynRealPart from SAP to PartImageNet [9], which has high-quality part segmentation annotation and provides extensive pose configurations.

To brigde the domain gap between synthetic and real, we test 3 state-of-the-art Syn-to-Real domain adaptation methods [16–18] used for semantic segmentation on SynRealPart, but fail to achieve decent results. Semantic animal part segmentation is more challenging than semantic segmentation tasks because semantic parts of animals often have similar appearance and highly varying shapes.

To address this challenge, we propose a simple yet effective method called Class-Balanced Fourier Data Mixing (CB-FDM) which consists of two parts. The first part Fourier Data Mixing (FDM) aligns the spectral amplitudes of synthetic and real images before mixing them for real domain training, thereby making the mixed images have more similar frequency content with real images. Specifically, we reconstruct the synthetic image with its original spectral phase and spectral amplitude of the real image. The reconstructed image is then mixed with the real image for training in real domain. Furthermore, we propose to use Class-Balanced Pseudo-Label Re-Weighting (CB) on certain minority class in terms of pixel frequency to alleviate the influence of the imbalanced class distribution in SAP.

We empirically evaluate the effectiveness of our method on the SynRealPart benchmark and achieve non-trivial improvement compared to various domain adaptation methods. Specifically, we improve DAformer [16] from 48.08 to 58.04 mIoU. Notably, our experiments also reveal that the learned parts from synthetic tiger and horse can be efficiently transferred (i.e. without using real labels) across all quadrupeds species in PartImageNet, even for species that have large shape variations with tiger and horse.

In summary, our main contributions are:

1. We construct a synthetic animal dataset of tigers and horses with larger diversity in pose space, named Synthetic Animal Parts (SAP), to facilitate research in animal part segmentation.

2. We set up a new Syn-to-Real benchmark of animal part segmentation from SAP to PartImageNet and propose a simple yet effective method CB-FDM to adapt Syn-to-Real methods designed for semantic segmentation to animal part segmentation.

3. We reveal that the learned parts from synthetic tigers and horses are transferable across all quadrupeds in PartImageNet, which supports that core set selection for each animal category could be an effective solution

to limited data, further underscoring the utility and potential applications of animal part segmentation.

2. Related Work

2.1. Part Segmentation

Segmenting object parts is a long-standing problem in computer vision and there is a rich literature on the topic. The pioneering work Pictorial Structure [19] along with following works [20–24] explicitly model parts and their spatial relations to the whole object. These methods share a common theme that the object-part models provide rich representations of objects and help interpretability. However, in the era of deep learning with data-driven models, research on part-based models gets hindered due to the lack of large-scale datasets. As a result, most recent works [25–31] mainly concentrate on unsupervised or self-supervised co-part segmentation. Both rigid [32, 33] and non-rigid objects [34–37] have been studied in part segmentation, but for non-rigid objects, recent works mainly focus on human part segmentation [34, 35, 38, 39] while there are still limited progress for animals due to the severer scarce of data. In this work, we propose a new direction to solve animal part segmentation by utilizing synthetic data, which is much cheaper and easier to obtain compared to the expensive real data. We further explore how to transfer the models from synthetic to real in an unsupervised manner and achieve promising results.

2.2. Synthetic Data

Synthetic data generated by computer graphics techniques are effective for model diagnosis [40, 41] and have boosted performance in many real-world application domains [10, 11, 42–46]. For synthetic animals, Haggag et al. [47] uses a marker-based motion-capture (MoCap) system to manually generate the animal poses which is time-consuming to generate more diverse poses. Mu et al. [10] uses 3D CAD animal models with their given animation sequences for data generation. However, the number of poses of CAD models is limited due to their animation sequences and thus is hard to scale up. We propose to use SMAL models [12] to generate synthetic data with more diverse poses as a supplementary for the CAD synthetic data.

2.3. Fourier Domain Bridging

In recent years, there has been a renewed interest in using Fourier transform based methodologies in efforts to solve problems like domain adaptation [48–50], domain generalization [51, 52], domain gap reduction [53], etc. Few works [48, 53] swap only low frequency component of the amplitude spectrum in order to learn better domain bridging features by aping target image style. Others [50, 52] employ Fourier amplitude information to generate synthetic or noisy adversarial images using source domain amplitude spectrum. There also have been attempts [49, 51] to preserve the phase information of an image to learn better domain bridging features - by creating images with interpolated amplitude spectrum [51] or mapping between the phase information of the source and target domains [49]. Our Fourier Cross-Domain Data Mixing does not employ selective spectrum swap like [48, 53] for whole images, regularize optimization using adversarial images [50, 52] or phase spectrum data [49, 51] or simply interpolate between domain spectrum [51]. We simply utilize Fourier domain information along with spatial data mixing to help the model learn better cross-domain features. Our method does take inspiration from the simple, yet effective aforementioned works regarding utilisation of properties of Fourier transform.

3. Synthetic Dataset

By evaluating the results of utilizing CAD data for Syn-to-Real animal part segmentation on PartImageNet, we identify a few failure cases involving unusual poses (e.g. lying) and animals with self-occlusions. One typical example is shown in Fig 2. Adding synthetic data with more unusual poses is the most intuitive solution. However, the animation sequences for each CAD animal model provide limited poses and public animal motion capture data is also scarce, which make the process difficult. Therefore, we opt to utilize SMAL models [12], which built a parametric way to represent the animal shape and pose based on strong prior. SMAL models are able to precisely reconstruct animal poses from 2D images by using the animal keypoints annotation and silhouette (i.e. foreground) mask. Moreover, people only need to annotate parts for one SMAL model as these models share the same vertex IDs, which makes the data generation super efficient. We present the details of our data generation process and statistics below.

Figure 2. **Failure cases when using CAD synthetic data only.** Due to the limited poses contained in CAD models, the model fails to segment the torso out and also predicts inaccurate boundaries.

3.1. Data Generation

(1) **2D Annotation.** Firstly, we select several natural animal images depicting poses not including in animation sequences

of CAD models from online resources. Secondly, we carefully annotate 26 keypoints per image based on the original keypoints definition of SMAL model. Inspired by [14], we replace the manual labeling process for silhouettes with the prediction of pre-trained object segmentation model [15].
(2) **SMAL Fitting.** With the keypoints and silhouette masks, we utilize SMALR [54] which recovers refined 3D models from 2D images. For detailed fitting process, we refer to [54] for more information.
(3) **3D Part Annotation.** The annotation of 3D parts is done by grouping the vertex IDs of each semantic part. Firstly, we select an SMAL model of arbitrary animal in a typical pose. Secondly, we use Blender [55] to group the vertex IDs of each part and save them. Since all SMAL models share the same vertex IDs, this part annotation can be directly applied to other SMAL models.
(4) **Rendering Images and Part Segmentation Masks.** Following previous work [10, 11], we use Blender as our render and randomize render parameters (e.g., viewpoint, lighting, and object texture) to promote domain generalization. The background images are randomly sampled from COCO [56]. The 2D part segmentation mask is obtained through directly projecting the annotated parts in the third step.

3.2. Dataset Statistics

We create our SMAL data utilizing the aforementioned pipeline. Specifically, we generate a total of 4,400 synthetic tiger images from 11 distinct poses. Each pose encompasses 100 viewpoints and 4 transformations, notably rotations. Similarly, we produce 2,000 synthetic horse images derived from 10 poses, with 100 viewpoints and 2 transformations. All animals are rendered with randomly selected textures from real images. It is important to highlight that our SMAL synthetic data offers a broad range of unusual poses, including lying, climbing, and other movements beyond walking and running, which are not included in the CAD synthetic data introduced in [10].

We further integrate our SMAL synthetic data with the existing CAD synthetic data [10], creating a comprehensive dataset specifically designed for animal part segmentation, named Synthetic Animal Parts (SAP). This combined dataset encompasses a total of 14,400 images for tigers and 12,000 images for horses. SAP offers diverse pose configurations for both tigers and horses, accompanied by accurate part masks. We believe that SAP will serve as valuable resources for advancing research in animal part segmentation.

4. Methods

In this section, we begin by presenting the formulation for syn-to-real part segmentation. Subsequently, we introduce the intuition and details of Fourier Data Mixing (FDM). Finally, we illustrate the motivation behind Class-Balanced Pseudo-Label Re-Weighting (CB).

4.1. Preliminaries

Syn-to-Real Part Segmentation Similar to semantic segmentation, a part segmentation model predicts the pixel-wise label for parts of a object where each part is a category. For example, in our quadruped animals setup, the part classes are head, torso, leg, tail of a quadruped animal. We denote the source domain as $\mathcal{D}_s = \{(x_s^{(i)}, y_s^{(i)})\}_{i=1}^{N_s}$ with N_s samples drawn from the synthetic domain, where $x_s^{(i)} \in X_s$ is an image, $y_s^{(i)} \in Y_s$ is the corresponding pixel-wise one-hot label over $K+1$ classes (including background). Note that K is the number of part classes. Similarly, the unlabeled target domain is denoted as $\mathcal{D}_t = \{x_t^{(i)}\}_{i=1}^{N_t}$ with N_t samples drawn from the real domain. This work aims to learn a part segmentation model that can effectively transfer part knowledge from the synthetic domain to the real domain. In addition, the part segmentation model is also assumed to have ability to transfer parts from one object class to a similar object class.

4.2. Fourier Data Mixing

Recent unsupervised syn-to-real translation methods [16–18, 57, 58] use self-training (i.e. using pseudo labelled real images for training on target domain) framework. The quality of pseudo-labels for the target images is crucial for achieving satisfactory convergence. As a solution, [59] proposed mixing the source and target domain image patches with binary masks obtained using various mixing algorithms [60, 61]. On the other hand, spectral information obtained through Fourier transform can be utilized to provide a global description of the image [62] as well as help in learning domain bridging features [49, 51]. Inspired by both of these concepts, we align the global information of the mixed regions to help the model learn better cross-domain features by aligning the spectral amplitudes of the mixed images. Let $\mathcal{F}^A$, $\mathcal{F}^P : \mathbb{R}^{H\times W\times 3} \rightarrow \mathbb{R}^{H\times W\times 3}$ be the amplitude and phase components of the Fourier transform $\mathbf{F}$ of an RGB image, i.e., for a single channel image x we have:

$$\mathcal{F}(x)(m,n) = \sum_{h,w} x(h,w)e^{-j2\pi(\frac{h}{H}m+\frac{w}{W}n)}, j^2 = -1 \quad (1)$$

, which can be implemented efficiently use the Fast Fourier Transform(FFT) algorithm [63]. $\mathcal{F}^{-1}$ is the inverse Fourier transform that maps the spectral signals back to the image space. Then in the mixed sampling stage, given two random samples $(x_s, y_s) \sim \mathcal{D}_s, x_t \sim \mathcal{D}_t$, we use FFT to get the spectral signals(amplitude and phase) from both, then the Fourier alignment can be formulated as:

$$x_{mixed} = \mathbf{M}\odot\mathcal{F}^{-1}(\mathcal{F}^A(x_t), \mathcal{F}^P(x_s)) + (\mathbf{1}-\mathbf{M})\odot x_t \quad (2)$$

$$y_{mixed} = \mathbf{M}\odot y_s + (\mathbf{1}-\mathbf{M})\odot \hat{y}_t \quad (3)$$

, where $\mathbf{M}$ denotes a binary mask generated by ClassMix [60], indicating which pixel needs to be copied from the augmented source domain and pasted to the target domain, $\mathbf{1}$ is a mask filled with ones, and $\odot$ represents the element-wise multiplication operation, and $\hat{y}_t$ is the pseudo-label of x_t. In the formulation, the amplitude of the source image $\mathcal{F}^A(x_s)$ is replaced by that of the target image x_t. Then the modified spectral representation of x_s, with its phase component unchanged, is mapped back the image domain to get the augmented image. We hypothesize that by making the source region of the mixed images have more similar frequency content with the target region, it will be harder for the model the learn the difference between the source and target regions and force the model to learn more domain-invariant features, thereby achieving better performance.

4.3. Class-Balanced Pseudo-Label Re-Weighting

Our synthetic dataset exhibiting a class imbalanced distribution in terms of pixel frequency 1. Although existing methods [16–18,57] employ the Rare Class Sampling (RCS) training strategy to sample the source images which contain minority classes in terms of pixel frequency more often, the supporting gradients [64] for some minority classes may still be very limited at the early training stage. It arises due to the fact that many images containing minority classes in SAP marginally surpass the pre-defined threshold of pixel number in RCS. To prevent RCS from sampling in a limited range of images, we are not able to set a high threshold which help the sampled images to have relatively higher pixel frequency for the minority class. Therefore, even one image containing certain minority class is sampled by RCS, the supporting gradients may still be limited. In the research of semi-supervised learning (SSL) in the context of class-imbalanced data for the classification task, there is an observation that the undesired performance of existing SSL algorithms on imbalanced data is mainly due to low recall on minority classes in terms of number of samples, but the precision on minority classes is surprisingly high [65]. We observe a similar phenomenon w.r.t animal head part, which is one of the minority classes in terms of pixel frequency in our synthetic data. The predictions of the animal head parts for real images hardly give false positive results but mainly the false negative. Therefore, in order to boost pseudo-label confidence for the head part, we give it more weight. We name the multiplicative factor of the pseudo-label weights of animal head as β for convenience. This technique mitigates the issue that the model may receive very limited supporting gradients for animal head class at the early stage of training, leading to unsatisfactory performance.

5. Experiments

In this section, we first provide our implementation details including how we construct train and test set on SynRealPart and training settings in Sec. 5.1. After setting the stage, we introduce our main results, compared with state-of-the-art methods in Sec. 5.2, followed by ablation studies in Sec. 5.3 to validate the key designs in our model. In the end, we further explore the part knowledge transfer in Sec. 5.4 that part segmentation results are transferable among species of similar structures regardless of shape and texture difference which points to a promising future direction. Qualitative visualization results are presented as well.

Table 1. **Pixel frequency of 4 classes in the Synthetic Animal Parts (SAP) dataset.** We compute the statistics of pixel frequency in SAP which exhibits the class-imbalance distribution.

class	head	torso	leg	tail
pixel frequency	12.7%	55.6%	25.9%	5.8%

5.1. Implementation Details

Data For synthetic data, we utilize 23520 images (11520 synthetic tiger images + 12000 synthetic horse images) in SAP for training. For real training data, we select 5942 quadrupeds images from PartImageNet [9] which excludes tiger images. Note that PartImageNet doesn't have horse images. Then we can get a UDA setting that all tiger and horse information are learned from our synthetic data. We select another 1213 quadrupeds images as our main test set which includes tiger images.

Training settings. We conduct our experiments using SegFormer [66], DAFormer [16], HRDA [17] and SePiCo [18]. MiT-b5 (Mix Transformer encoders) pre-trained on ImageNet-1k is adopted as the backbone for above methods. If not specified, we train all our models with batch size 2 on a single GPU for 30k iterations. We set the learning rate of the decoder head to be 6e-4 and the backbone has a learning rate multiplier 0.1. We use AdamW [67] optimizer with weight decay 0.01. For data augmentation, we adopt random color jittering. The input image is cropped into 512×512 (1024×1024 for HRDA).

5.2. Main Results

Table 2 summaries our results on our main test set. SegFormer [66] supervisedly trained on SAP achieves 42.21 mIoU on the test set. We observe that naively applying state-of-the-art semantic segmentation syn-to-real methods sometimes can not bring significant improvement (i.e. from 42.21 to 43.87 in terms of mIoU for HRDA [17]). Notably, after applying our proposed Class-Balanced Fourier Data Mixing (CB-FDM), all syn-to-real methods get non-trivial improvement. For DAFormer, the improvement gain is up to 9.96 mIoU while we also improve HRDA by 4.47 and SePiCo [18] by 2.27 in terms of mIoU. SePiCo [18] which combines DAFormer [16] with contrastive learning shows the optimal performance in direct application but not very

Table 2. **Comparison of part segmentation (mIoU) on our test set.** CB-FDM brings non-traivial improvements to all the syn-to-real methods, especially for DAFormer [16], which makes DAFormer + CB-FDM the optimal syn-to-real solution in our current benchmark in terms of mIoU. Numbers are averaged over 3 random seeds. *: train SegFormer with synthetic data and real data successively for each iteration (equivalent to replacing pseudo-labels with real labels in DAFormer). "bg" stands for background.

data	method	head	torso	leg	tail	bg	mIoU
SAP	SegFormer [66]	49.71	38.84	31.86	7.53	83.11	42.21
PartImageNet [9]		85.79	72.69	60.00	58.27	96.81	74.71
SAP + Unlabeled PartImageNet	DAFormer [16]	44.94	48.66	44.11	12.21	90.45	48.08
	DAFormer + CB-FDM	69.40	56.81	49.17	21.42	93.39	58.04
	HRDA [17]	44.91	37.16	36.51	12.75	88.00	43.87
	HRDA + CB-FDM	64.21	40.69	42.03	6.54	88.21	48.34
	SePiCo [18]	55.67	57.83	42.53	15.6	91.29	52.58
	SePiCo + CB-FDM	58.72	52.71	49.37	21.17	92.27	54.85
SAP + PartImageNet	SegFormer*	86.33	73.37	62.30	56.65	96.72	75.08

Table 3. **CB-FDM ablation studies on our test set in terms of mIoU.** FDM & CB both provide non-trivial improvement based on DAFormer. Numbers are averaged over 3 random seeds.

method	FDM	CB	head	torso	leg	tail	bg	mIoU
DAFormer [16]	✗	✗	44.94	48.66	44.11	12.21	90.45	48.08
	✓	✗	50.14	50.72	49.17	20.74	93.11	52.77
	✗	✓	71.69	56.21	46.37	16.15	92.19	56.52
	✓	✓	69.40	56.81	49.17	21.42	93.39	**58.04**

sensitive to CB-FDM. We hypothesis that it is because CB-FDM is designed for the cross-entropy losses of the mixed images while SePiCo additionally has one contrastive loss for source images with their groundtruth and one contrastive loss for real images with their pesudo-labels, which mitigate the influence of CB-FDM. The syn-to-real results also reveal that part knowledge can be efficiently transferred among objects with similar structures regardless of shape and texture difference. With unlabeled real data and proper algorithms, the part knowledge of only synthetic tiger and horse can be much better adapted to all 46 quadrupeds in PartImageNet (i.e. DAFormer + CB-FDM improves the performance by 15.83 mIoU compared to training on SAP). We believe this finding can motivate the exploration in animal part segmentation since people will only need to select a core set of animals species in each animal category for training. In addtion, we supervisedly train a SegFormer on real training data from PartImageNet [9] and achieve 74.71 mIoU. When we supervisedly train on both synthetic and real data, there is another 0.37 improvement in terms of mIoU.

Visualizations. We also conduct qualitative comparison as illustrated in Fig. 3. We show that models trained only on SAP usually fail to generalize to real images while naively adapting semantic segmentation syn-to-real methods yields many incorrect part predictions. On the contrary, our CB-FDM with DAFormer is able to predict more accurate boundary for different parts even in the scenarios of challenging poses (e.g. row 2&3&4) and unseen species that have large shape difference with tiger and horse (e.g. row 5&6).

5.3. Ablation Study

Effectiveness of proposed modules. Table 3 summarizes the effects of the key designs in our method. We note that after applying Fourier Data Mixing (FDM), we can obtain a general improvement on all classes which lead to an overall improvement of 4.69 on mIoU. Class-Balanced (CB) sampling brings a significant improvement on head class (i.e. 26.75 mIoU) as it pays more attention to it while also improves the performance on all the other classes. When combining these two parts together, we obtain our final method Class-Balanced Fourier Data Mixing (CB-FDM) and achieves 58.04 mIoU. We can notice slight performance drop on head but improvements on all the other classes compared to using CB only. We assume that FDM prevents the over-fitting on head when using CB and thus is a good combination with it.

Influence of balance strength β. Tab. 4 presents our results on controlling the balance strength through β. As can be observed, a reasonable strong balance weight (i.e. $\beta > 1.5$) is required to achieve good results, while setting it too large will also harm the performance as well. We set β to be 2 as our default setting according to this experimental results.

Synthetic Data Source. Tab. 5 shows ablation studies on using our SMAL synthetic data. As we can observe from the comparisons, after introducing our SMAL synthetic data, SegFormer(i.e. synthetic only) achieves 1.6 improvement while DAFormer + CB-FDM achieves 2.59 improvement in terms of mIoU. However, we notice a performance drop on

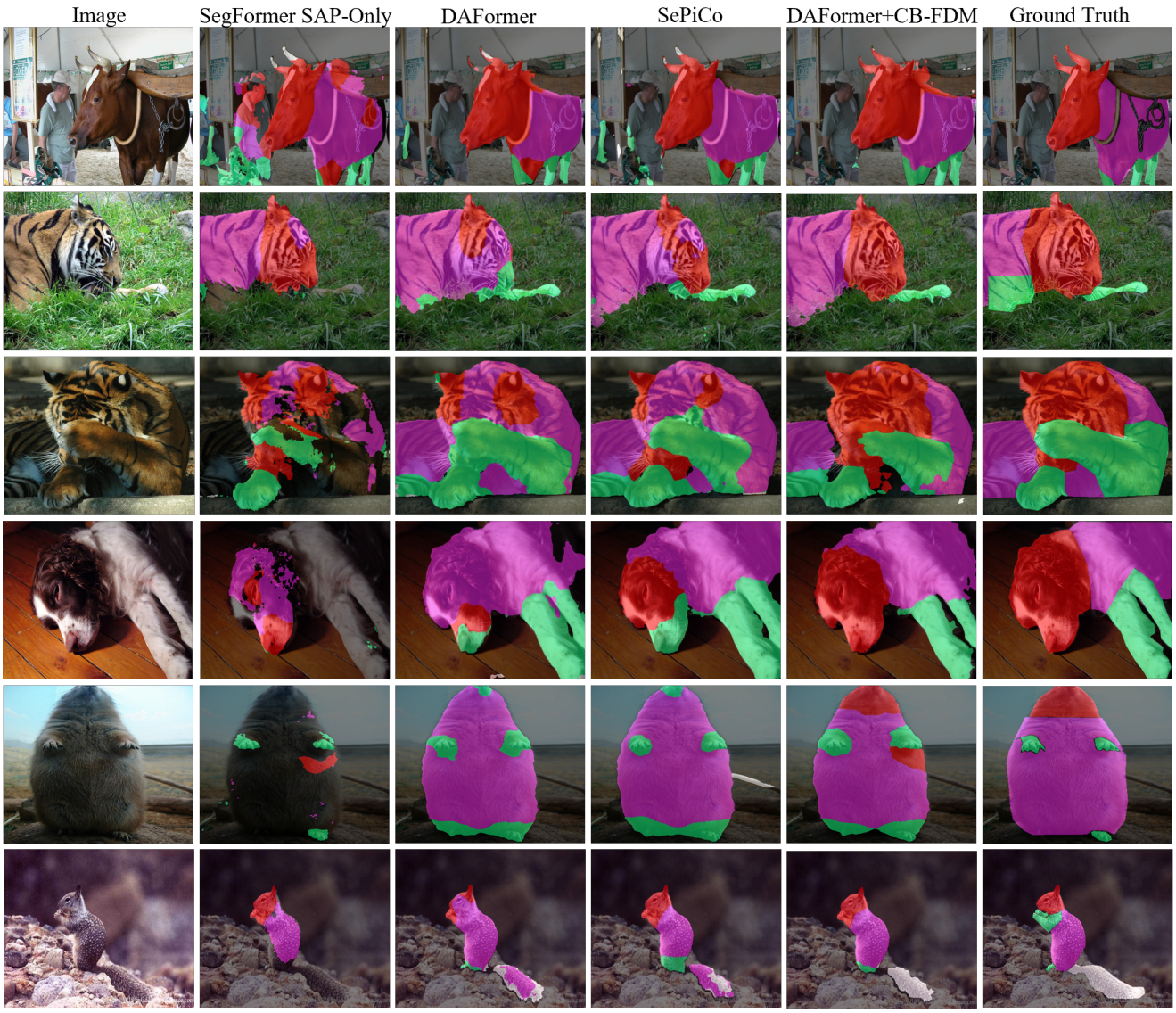

Figure 3. **Qualitative comparisons for different methods on our test set.** Note that our CB-FDM with DAFormer [16] produces more accurate part segmentation results in challenging poses (e.g., row 2&3&4) and unseen species that have large shape difference with tiger and horse (e.g., row 5&6).

Table 4. **Parameter study of the class balance weight β.** We achieve the optimal performance when $\beta = 2$. Numbers are averaged over 3 random seeds.

method	β	head	torso	leg	tail	bg	mIoU
DAFormer [16] + CB	1.5	58.97	52.89	41.87	19.54	92.17	53.09
	2	71.69	56.21	46.37	16.15	92.19	**56.52**
	2.5	71.72	57.74	45.86	13.67	92.46	56.29
	3	72.68	59.86	45.79	8.32	92.04	55.74

tail after introducing SMAL data. We hypothesis that one potential reason is the inaccurate tail shape for SMAL fitting algorithms when tails are self-occluded for unusual poses in real images.

5.4. Zero-Shot Part Knowledge Transfer

In Sec. 5.2, we already discussed that Table 2 implies the part knowledge from synthetic tiger and horse can be

Table 5. **Ablation study for adding SMAL synthetic data.** Introducing SMAL data lead to a non-trivial improvement. Numbers are averaged over 3 random seeds.

Method	synthetic data source		head	torso	leg	tail	bg	mIoU
	CAD	*SMAL*						
SegFormer [66]	✓	✗	44.51	35.72	31.45	8.20	83.19	40.61
	✓	✓	49.71	38.84	31.86	7.53	83.11	**42.21**
DAFormer [16]	✓	✗	59.09	45.40	48.88	31.98	91.90	55.45
+ CB-FDM	✓	✓	69.40	56.81	49.17	21.42	93.39	**58.04**

Table 6. **Part segmentation results (mIoU) of *Horse* and *Tiger* settings.** *Horse* setting is trained with synthetic horse and horse-like real images, and then test on tiger-like real images. Similarly, *Tiger* setting is trained with synthetic tiger and tiger-like real images, and then test on horse-like real images. Numbers are averaged over 3 random seeds.

settings	method	head	torso	leg	tail	bg	mIoU
Tiger	DAFormer + CB-FDM	32.97	57.52	46.46	2.14	94.10	46.64
Horse		71.22	54.63	38.26	12.07	90.11	53.26

efficiently transferred to all quadrupeds in PartImageNet. To further explore the power of this transfer ability, we design 2 more zero-shot settings for only tiger and horse respectively because they have relatively large shape difference. Since PartImageNet doesn't have horse classes, we extend the standard to tiger-like and horse-like animals. We select a tiger-like (i.e. tiger, cheetah, lion) animal set which includes 630 images and a horse-like (i.e. goat, deer, buffalo, etc) animal set which includes 1016 images from PartImageNet. Then we have the following 2 unsupervised syn-to-real settings: 1) *Tiger*: Train on synthetic tiger and tiger-like set, and test on horse-like set; 2) *Horse*: Train on synthetic horse and horse-like set, and test on tiger-like set. From table 5.4, we can see both settings can transfer the torso knowledge pretty good as that is the part which have the most similar shape. We assume the huge difference in head and tail performance between these 2 settings is mainly caused by the ambiguity problem of tail and horn (refer to supplementary for its visualization). Horn is an unseen parts for our synthetic data. Our real groundtruth regard it as a part of head while the model trained on synthetic often predicts it as tail. Horse-like set mainly consists of animals with horns, which leads to a huge performance drop on head and tail for *Tiger* setting. Without the ambiguity issue, we can see the *Horse* setting has achieved 53.26 mIoU even with a tiny amount of unlabeled real data and this zero-shot setup. We believe these findings point to a promising future direction in solving limited data problems in animal part segmentation and constructing training data more efficiently (i.e. core set selection for each animal category).

6. Conclusion

In this paper, we propose to use SMAL models for efficiently generating synthetic data with diverse pose configurations and further construct a synthetic animal dataset of tigers and horses with part segmentation groundtruth termed as Synthetic Animal Parts (SAP). Then we set up a new Syn-to-Real benchmark of animal part segmentation from SAP to PartImageNet called SynRealPart, and we propose a simple yet effective method called Class-Balanced Fourier Data Mixing (CB-FDM) consisting of Fourier Data Mixing (FDM) and Class-Balanced Pseudo-label Re-weighting (CB) to improve the performance of existing unsupervised syn-to-real adaptation methods designed for semantic segmentation on it. Our experiments also reveal that the learned parts from synthetic tiger and horse are transferable across all quadrupeds in PartImageNet, which supports that core set selection for each animal category could be an effective solution to limited data, further underscoring the utility and potential applications of animal part segmentation.

Limitations. At present, our SAP dataset focuses exclusively on two animal species: tigers and horses, which are two representative animals for quadrupeds. For quadrupeds, we plan to add one or two quadrupeds which have horns to solve the ambiguity problems between horn and tail. We also plan to expand our synthetic animal data to contain more animal categories like bird and reptile. While SMAL models are for quadrupeds only, we need to explore other efficient solutions to creating diverse poses, but using animation sequences of CAD models for normal poses and using 3D models reconstructed from real images for unusual poses may still be our core strategy. Furthermore, it is important to note that Class-Balanced Pseudo-Label Re-Weighting fails in "tail", resulting in satisfactory performance only for the "head" class. One main reason is the ambiguity problem between tail and horn which may be solved by adding synthetic animals with horns. However, tail is still the hardest part to segment since it has the biggest shape and deformation variance among different animals. Improving segmentation accuracy on tail is quite challenging and remains unsolved. Lastly, while our model demonstrates the ability to transfer part knowledge across different animal species, it still lacks the capability to handle unseen parts (i.e. horns).

References

[1] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR 2011*, pages 1385–1392, 2011. 1

[2] Jian Dong, Qiang Chen, Xiaohui Shen, Jianchao Yang, and Shuicheng Yan. Towards unified human parsing and pose estimation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 843–850, 2014. 1

[3] Hossein Azizpour and Ivan Laptev. Object detection using strongly-supervised deformable part models. In Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I*, volume 7572 of *Lecture Notes in Computer Science*, pages 836–849. Springer, 2012. 1

[4] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan L. Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. *CoRR*, abs/1406.2031, 2014. 1

[5] S. Eslami and Christopher Williams. A generative model for parts-based object segmentation. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 1

[6] Peng Wang, Xiaohui Shen, Zhe L. Lin, Scott Cohen, Brian L. Price, and Alan L. Yuille. Joint object and part segmentation using deep learned potentials. *CoRR*, abs/1505.00276, 2015. 1

[7] Ning Zhang, Jeff Donahue, Ross B. Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. *CoRR*, abs/1407.3867, 2014. 1

[8] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan L. Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. *CoRR*, abs/1406.2031, 2014. 1

[9] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts, 2022. 1, 2, 5, 6

[10] Jiteng Mu, Weichao Qiu, Gregory Hager, and Alan Yuille. Learning from synthetic animals, 2020. 1, 3, 4

[11] Qing Liu, Adam Kortylewski, Zhishuai Zhang, Zizhang Li, Mengqi Guo, Qihao Liu, Xiaoding Yuan, Jiteng Mu, Weichao Qiu, and Alan Yuille. Learning part segmentation through unsupervised domain adaptation from synthetic vehicles, 2022. 1, 3, 4

[12] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3d menagerie: Modeling the 3d shape and pose of animals, 2017. 1, 3

[13] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015. 1

[14] Benjamin Biggs, Thomas Roddick, Andrew W. Fitzgibbon, and Roberto Cipolla. Creatures great and SMAL: recovering the shape and motion of animals from video. *CoRR*, abs/1811.05804, 2018. 1, 4

[15] Feng Li, Hao Zhang, Huaizhe xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation, 2022. 1, 4

[16] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation, 2022. 2, 4, 5, 6, 7, 8

[17] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation, 2022. 2, 4, 5, 6

[18] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation, 2023. 2, 4, 5, 6

[19] Martin A Fischler and Robert A Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on computers*, 100(1):67–92, 1973. 3

[20] Markus Weber, Max Welling, and Pietro Perona. Unsupervised learning of models for recognition. In *European conference on computer vision*, pages 18–32. Springer, 2000. 3

[21] Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. *International journal of computer vision*, 61(1):55–79, 2005. 3

[22] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006. 3

[23] Song-Chun Zhu and David Mumford. *A stochastic grammar of images*. Now Publishers Inc, 2007. 3

[24] Ross Girshick, Pedro Felzenszwalb, and David McAllester. Object detection with grammar models. *Advances in neural information processing systems*, 24, 2011. 3

[25] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. Scops: Self-supervised co-part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 869–878, 2019. 3

[26] Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised part discovery from contrastive reconstruction. *Advances in Neural Information Processing Systems*, 34:28104–28118, 2021. 3

[27] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Unsupervised part segmentation through disentangling appearance and shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8355–8364, 2021. 3

[28] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. Repurposing gans for one-shot semantic part segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4475–4485, 2021. 3

[29] Samir Yitzhak Gadre, Kiana Ehsani, and Shuran Song. Act the part: Learning interaction strategies for articulated object part discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15752–15761, 2021. 3

[30] Adrian Ziegler and Yuki M Asano. Self-supervised learning of object parts for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14502–14511, 2022. 3

[31] Oindrila Saha, Zezhou Cheng, and Subhransu Maji. Improving few-shot part segmentation using coarse supervision. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 283–299. Springer, 2022. 3

[32] Wenhao Lu, Xiaochen Lian, and Alan Yuille. Parsing semantic parts of cars using graphical models and segment appearance consistency, 2014. 3

[33] Yafei Song, Xiaowu Chen, Jia Li, and Qinping Zhao. Embedding 3d geometric features for rigid object part segmentation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 580–588, 2017. 3

[34] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network, 2018. 3

[35] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. Deep human parsing with active template regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(12):2402–2414, dec 2015. 3

[36] Jianyu Wang and Alan Yuille. Semantic part segmentation using compositional model combining shape and appearance, 2014. 3

[37] Kota Yamaguchi, M. Hadi Kiapour, Luis E. Ortiz, and Tamara L. Berg. Parsing clothing in fashion photographs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3570–3577, 2012. 3

[38] Wenguan Wang, Zhijie Zhang, Siyuan Qi, Jianbing Shen, Yanwei Pang, and Ling Shao. Learning compositional neural information fusion for human parsing, 2020. 3

[39] Si Liu, Yao Sun, Defa Zhu, Guanghui Ren, Yu Chen, Jiashi Feng, and Jizhong Han. Cross-domain human parsing via adversarial feature and label adaptation, 2018. 3

[40] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *CoRR*, abs/1612.06890, 2016. 3

[41] Yi Zhang, Weichao Qiu, Qi Chen, Xiaolin Hu, and Alan L. Yuille. Unrealstereo: A synthetic dataset for analyzing stereo vision. *CoRR*, abs/1612.04647, 2016. 3

[42] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. *CoRR*, abs/1504.06852, 2015. 3

[43] Ankur Handa, Viorica Patraucean, Vijay Badrinarayanan, Simon Stent, and Roberto Cipolla. Scenenet: Understanding real world indoor scenes with synthetic data. *CoRR*, abs/1511.07041, 2015. 3

[44] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and

Thomas Vetter. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2261–2268, 2019. 3

[45] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. *CoRR*, abs/1804.06516, 2018. 3

[46] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. *CoRR*, abs/1701.01370, 2017. 3

[47] Hussein Haggag, Ahmed Abobakr, Mohammed Hossny, and Saeid Nahavandi. Semantic body parts segmentation for quadrupedal animals. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 000855–000860, 2016. 3

[48] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation, 2020. 3

[49] Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. Phase consistent ecological domain adaptation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9008–9017, 2020. 3, 4

[50] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Rda: Robust domain adaptation via fourier adversarial attacking. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8968–8979, 2021. 3

[51] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14378–14387, 2021. 3, 4

[52] Minyoung Kim, Da Li, and Timothy Hospedales. Domain generalisation via domain adaptation: An adversarial fourier amplitude approach. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[53] V. Kumar, S. Srivastava, R. Lal, and A. Chakraborty. Caft: Class aware frequency transform for reducing domain gap. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2525–2534, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society. 3

[54] Silvia Zuffi, Angjoo Kanazawa, and Michael J. Black. Lions and tigers and bears: Capturing non-rigid, 3D, articulated shape from images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3955–3963. IEEE Computer Society, 2018. 4

[55] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 4

[56] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 4

[57] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation, 2023. 4, 5

[58] Lin Chen, Zhixiang Wei, Xin Jin, Huaian Chen, Miao Zheng, Kai Chen, and Yi Jin. Deliberated domain bridging for domain adaptive semantic segmentation, 2022. 4

[59] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. DACS: domain adaptation via cross-domain mixed sampling. *CoRR*, abs/2007.08702, 2020. 4

[60] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning, 2020. 4, 5

[61] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. *CoRR*, abs/1905.04899, 2019. 4

[62] Olivier Joubert, Guillaume Rousselet, Michèle Fabre-Thorpe, and Denis Fize. Rapid visual categorization of natural scene contexts with equalized amplitude spectrum and increasing phase noise. *Journal of vision*, 9:2.1–16, 02 2009. 4

[63] M. Frigo and S.G. Johnson. Fftw: an adaptive software architecture for the fft. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 3, pages 1381–1384 vol.3, 1998. 4

[64] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey, 2023. 5

[65] Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning, 2021. 5

[66] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *CoRR*, abs/2105.15203, 2021. 5, 6, 8

[67] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. 5