PASTA: Proportional Amplitude Spectrum Training Augmentation for Syn-to-Real Domain Generalization

Prithvijit Chattopadhyay* Kartik Sarangmath* Vivek Vijaykumar Judy Hoffman
Georgia Institute of Technology
{prithvijit3, ksarangmath3, vivekvjk, judy}@gatech.edu

https://github.com/prithv1/PASTA

Abstract

*Synthetic data offers the promise of cheap and bountiful training data for settings where labeled real-world data is scarce. However, models trained on synthetic data significantly underperform when evaluated on real-world data. In this paper, we propose Proportional Amplitude Spectrum Training Augmentation (*PASTA*), a simple and effective augmentation strategy to improve out-of-the-box synthetic-to-real (syn-to-real) generalization performance.* PASTA *perturbs the amplitude spectra of synthetic images in the Fourier domain to generate augmented views. Specifically, with* PASTA *we propose a structured perturbation strategy where high-frequency components are perturbed relatively more than the low-frequency ones. For the tasks of semantic segmentation (GTAV→Real), object detection (Sim10K→Real), and object recognition (VisDA-C Syn→Real), across a total of 5 syn-to-real shifts, we find that* PASTA *outperforms more complex state-of-the-art generalization methods while being complementary to the same.*

1. Introduction

For complex tasks, deep models often rely on training with substantial labeled data. Real-world data can be expensive to label and an available labeled training set often captures only a limited set of real-world appearance diversity. Synthetic data offers an opportunity to cheaply generate diverse samples that can better cover the anticipated variance of real-world test data. However, models trained on synthetic data often struggle to generalize to real world data – e.g., the performance of a vanilla DeepLabv3+ [13] (ResNet-50 backbone) architecture on semantic segmentation drops from 73.45 mIoU on GTAV [58] to 28.95 mIoU on Cityscapes [18] for the same set of classes. Several approaches have been considered to tackle this problem.

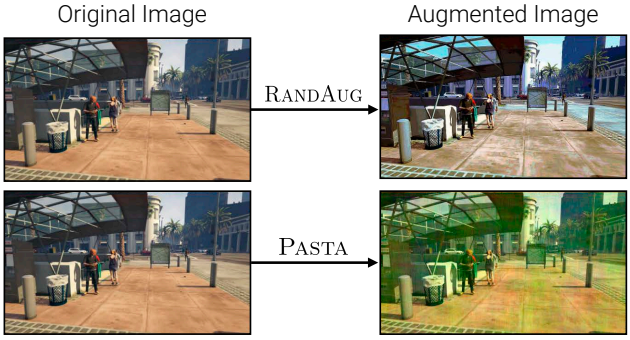

Figure 1: **PASTA augmentation samples.** A synthetic image from GTAV [58] when augmented with RandAugment [19] **[Top]** and PASTA **[Bottom]**. We find that PASTA creates augmented views different from existing photometric operations.

In this paper, we propose a novel augmentation strategy, called Proportional Amplitude Spectrum Training Augmentation (PASTA), to address the synthetic-to-real generalization problem. PASTA, as an augmentation strategy for synthetic data, aims to satisfy three key criteria: (1) *strong out-of-the-box generalization performance*, (2) *plug-and-play* compatibility with existing methods, and (3) benefits across *tasks, backbones,* and *shifts*. PASTA achieves this by perturbing the amplitude spectra (obtained by applying 2D FFT to input images) of the source synthetic images in the Fourier domain. While prior work has explored augmenting images in the Fourier domain [71, 73, 31], they mostly rely on the observations that – (1) among the amplitude and phase spectra, phase tends to capture more high-level semantics [52, 51, 55, 25, 72] and (2) low-frequency (LF) bands of the amplitude spectrum tend to capture style information / low-level statistics (illumination, lighting, etc.) [73].

We further observe that synthetic images have less diversity in the high-frequency (HF) bands of their amplitude spectra compared to real images (see Sec. 3.3 for a detailed discussion). Motivated by these key observations, PASTA provides a structured way to perturb the amplitude spectra of

*Equal Contribution. Correspondence to prithviijt3@gatech.edu

source synthetic images to ensure that a model is exposed to more variations in high-frequency components during training. We empirically observe that by relying on such a simple set of motivating observations, PASTA leads to significant improvements in synthetic-to-real generalization performance – e.g., out-of-the-box GTAV [58]→Cityscapes [18] generalization performance of a vanilla DeepLabv3+ (ResNet-50 backbone) model improves from 28.95 mIoU to 44.12 mIoU – a 15+ absolute mIoU point improvement!

PASTA involves the following steps. Given an input image, we apply 2D Fast Fourier Transform (FFT) to obtain the corresponding amplitude and phase spectra in the Fourier domain. For every spatial frequency $[m, n]$ in the amplitude spectrum, we sample a multiplicative jitter value ϵ such that the perturbation strength increases monotonically with $[m, n]$, thereby, ensuring that higher frequency components in the amplitude spectrum are perturbed more compared to the lower frequency ones. Finally, given the perturbed amplitude and the original phase spectra, we can apply an inverse 2D Fast Fourier Transform (iFFT) to obtain the augmented image. This simple strategy of applying fine-grained structured perturbations to the amplitude spectra of synthetic images leads to strong out-of-the-box generalization without the need for specialized components, task-specific design, or changes to learning rules. Fig. 1 shows an example image augmented by PASTA.

To summarize, we make the following contributions:

- We introduce Proportional Amplitude Spectrum Training Augmentation (PASTA), a simple and effective augmentation strategy for synthetic-to-real generalization. PASTA perturbs the amplitude spectra of synthetic images so as to expose a model to more high-frequency variations.
- We show that applying PASTA sets the new state of the art for synthetic-to-real generalization for 3 tasks – Semantic Segmentation (GTAV [58]→{Cityscapes [18], Mapillary [48], BDD100K [76]}), Object Detection (Sim10K [33]→Cityscapes) and Object Recognition (VisDA-C [54] Syn→Real) – covering a total of 5 syn-to-real shifts with multiple backbones.
- Our experimental results demonstrate that PASTA– (1) frequently enables a baseline model to outperform previous state-of-the-art approaches that rely on specialized architectural components, additional synthetic or real data, or alternate objectives; (2) is complementary to existing methods; (3) outperforms prior adaptive object detection methods; and (4) either outperforms or is competitive with current augmentation strategies.

2. Related work

Domain Generalization (DG). DG typically involves training models on single or multiple labeled data sources to generalize well to novel test time data sources (unseen during training). Several approaches have been proposed to tackle domain generalization [4, 47], such as decomposing a model into domain invariant and specific components and utilizing the former to make predictions [23, 35], learning domain specific masks for generalization [8], using meta-learning to train a robust model [41, 65, 3, 12, 21], manipulating feature statistics to augment training data [79, 43, 49], and using models crafted based on risk minimization formalisms [2]. More recently, properly tuned ERMs (Empirical Risk Minimization) have proven to be a competitive DG approach [24], with follow-up work adopting various optimization and regularization techniques [62, 7] on top.

Single Domain Generalization (SDG). Unlike DG which leverages diversity across multiple sources for better generalization, SDG considers generalizing from a single source. Notable approaches for SDG use meta-learning [56] by considering strongly augmented source images as meta-target data (by exposing the model to increasingly distinct augmented views of the source data [68, 42]) and learning feature normalization schemes with auxiliary objectives [22].

Synthetic-to-Real Generalization (Syn-to-Real). Prior work on syn-to-real generalization has mostly focused on some specific methods, including learning feature normalization / whitening schemes [53, 17], using external data for style injection [36, 38], explicitly optimizing for robustness [15], leveraging strong augmentations / domain randomization [77, 38], consistency objectives [78] and using contrastive techniques to aid generalization [14]. Some approaches have also considered adapting from synthetic to real images, using techniques such as adversarial training [16], adversarial alignment losses [60], balancing transferability and discriminability [9] and feature alignment [64]. PASTA is more similar to the kind of methods adopting augmentations for improving out-of-the-box generalization We consider 3 of the most commonly studied syn-to-real generalization settings – (1) Semantic Segmentation - GTAV [58]→Real, (2) Object Detection - Sim10K [33]→Real and (3) Object Recognition - VisDA-C [54] Syn→Real. [46] recently proposed tailoring synthetic data for better generalization.

Fourier Generalization & Adaptation Methods. Prior work that explored augmenting images in the Fourier domain (as opposed to the pixel space) rely on a key empirical observation [52, 51, 55, 25] that the phase component of the Fourier spectrum tends to preserve high-level semantics, and therefore, they focused mostly on perturbing the amplitude. PASTA is in line with this style of approach. Amplitude Jitter (AJ) [71] and Amplitude Mixup (AM) [71] are methods similar to PASTA that augment images by perturbing their amplitude spectra. While AM mixes the amplitude spectra of different images, AJ applies uniform perturbation with a single jitter value ϵ. FSDR [31], on the other hand, isolates domain variant and invariant frequency components using extra data and sets up a learning paradigm. Building on top of [71], [74] adds a significance mask when linearly in-

terpolating amplitudes. [32] only perturbs image-frequency components that capture little semantic information. [66] uses an encoder-decoder to obtain high/low frequency features and augments images by adding noise to high frequency phase and low frequency amplitude. [75, 10] study how amplitude and phase perturbations impact robustness to natural corruptions [27]. In contrast to these works, PASTA's simple strategy of perturbing amplitude spectra in a structured way and leads to strong out-of-the-box generalization without the need for specialized components, extra data, task-specific design, or changes to learning rules.

3. Method

We investigate how well models trained on a single labeled synthetic source dataset generalize to real target data, without any access to target data during training.

3.1. Preliminaries: Fourier Transform

PASTA creates augmented views by applying perturbations to the Fourier amplitude spectrum. The amplitude $\mathcal{A}(x)$ and phase $\mathcal{P}(x)$ components of Fourier spectra of images have been widely used in image processing for several applications – for studying properties (e.g., periodic interefferences), compact representations (e.g, JPEG compression), digital filtering, etc – and more recently for generalizing and adapting deep networks by perturbing the amplitude spectra [71, 73]. We now cover preliminaries explaining how to obtain amplitude and phase spectra from an image.

Consider a single-channel image $x \in \mathbb{R}^{H \times W}$. The Fourier transform $\mathcal{F}(x)$ of x can be expressed as,

$$\mathcal{F}(x)[m,n] = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x[h,w] \exp\left(-2\pi i \left(\frac{h}{H}m + \frac{w}{W}n\right)\right) \tag{1}$$

where $i^2 = -1$ and m, n denote spatial frequencies.

The inverse Fourier transform, $\mathcal{F}^{-1}(\cdot)$, that maps signals from the frequency domain to the image domain can be defined accordingly. Note that the Fourier spectrum $\mathcal{F}(x) \in \mathbb{C}^{H \times W}$. If $\mathbf{Re}(\mathcal{F}(x)[\cdot,\cdot])$ and $\mathbf{Im}(\mathcal{F}(x)[\cdot,\cdot])$ denote the real and imaginary parts of the Fourier spectrum, the corresponding amplitude ($\mathcal{A}(x)[\cdot,\cdot]$) and phase ($\mathcal{P}(x)[\cdot,\cdot]$) spectra can be expressed as,

$$\mathcal{A}(x)[m,n] = \sqrt{\mathbf{Re}(\mathcal{F}(x)[m,n])^2 + \mathbf{Im}(\mathcal{F}(x)[m,n])^2} \tag{2}$$

$$\mathcal{P}(x)[m,n] = \texttt{arctan}\left(\frac{\mathbf{Im}(\mathcal{F}(x)[m,n])}{\mathbf{Re}(\mathcal{F}(x)[m,n])}\right) \tag{3}$$

Without loss of generality, we will assume for the rest of this section that the amplitude and phase spectra are zero-centered, *i.e.*, the low-frequency components (low $[m,n]$) have been shifted to the center (lowest frequency component is at the center). The Fourier transform and its inverse can be calculated efficiently using the Fast Fourier Transform (FFT) [50] algorithm. For an RGB image, we can obtain the Fourier spectrum (and $\mathcal{A}(x)[\cdot,\cdot]$ and $\mathcal{P}(x)[\cdot,\cdot]$) independently for each channel. For the following sections, although we illustrate PASTA using a single-channel image, it can be easily extended to multi-channel (RGB) images by treating each channel independently.

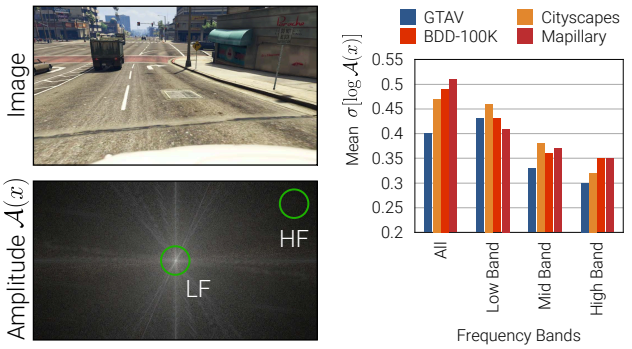

Figure 2: **Amplitude spectrum characteristics. [Left]** Sample amplitude spectrum (LF = Low Frequency, HF = High Frequency) for a single channel of a synthetic image from GTAV [58]. Note that the amplitude spectrums tend to follow a specific pattern – (for natural images) amplitude values tend to follow an inverse power law w.r.t. the spatial frequency [5, 63], *i.e.*, roughly, the amplitude at frequency f, $\mathcal{A}(f) \propto \frac{1}{f^\gamma}$, for some γ determined empirically. **[Right]** Variations in amplitude values across images. Synthetic images have less variance in high-frequency components of the amplitude spectra compared to real images.

3.2. Amplitude Spectrum Characteristics

Prior Observations. We first note that for natural images, the amplitude spectra $\mathcal{A}(x)$ has a specific structure – amplitude values tend to follow an inverse power law w.r.t. the spatial frequency [5, 63], *i.e.*, roughly, the amplitude at frequency f, $\mathcal{A}(f) \propto \frac{1}{f^\gamma}$, for some γ determined empirically (see Fig. 2 [Left]). Moreover, as noted earlier, a considerable body of prior work [52, 51, 55, 25] has shown that the phase component of the Fourier spectrum tends to preserve the semantics of the input image,[1] and the low-frequency (LF) components of the amplitude spectra tend to capture low-level photometric properties (illumination, etc.) [73]. Based on these observations, several methods [73, 31, 71] generate augmented views by modifying only the amplitude spectra of input images, leaving the phase information "unchanged" – by either copying the amplitude spectra from an image from another domain [73] or by introducing naive

[1] More accurately, small variations in the phase component can significantly alter the semantics of the image.

uniform perturbations [71]. PASTA introduces a fine-grained perturbation scheme for the amplitude spectra based on an additional empirical observation when comparing synthetic and real images.

Our Observation. Across a set of synthetic source datasets, we make the important observation that synthetic images tend to have smaller variations in the high-frequency (HF) components of their amplitude spectrum than real images.[2] Fig. 2 [Right] shows the standard deviation of amplitude values for different frequency bands per-dataset with the three real datasets (for high-band of Cityscapes [18], BDD-100K [76], Mapillary [48]) being significantly larger than the synthetic dataset (high band GTAV [61]). In appendix, we show how this phenomenon is consistent across (1) several syn-to-real shifts and (2) fine-grained frequency band discretizations. This phenomenon is likely a consequence of how synthetic images are rendered. For instance, in VisDA-C [54], the synthetic images are viewpoint images of 3D object models (under different lighting conditions), so it is unlikely for them to be diverse in high-frequency details. For images from GTAV [58], synthetic renderings can lead to contributing factors such as low texture variations – for instance, "roads" (one of the head classes in semantic segmentation) in synthetic images likely have less high-frequency variation compared to real roads.[3] Consequentially, to generalize well to real data, we would like to ensure that our augmentation strategy exposes the model to more variations in the high-frequency components of the amplitude spectrum during training. Exposing a model to such variations during training allows it to be invariant to this "nuisance" characteristic – which prevents overfitting to a specific syn-to-real shift.

3.3. Proportional Amplitude Spectrum Training Augmentation (PASTA)

Following the observation that synthetic data has lower amplitude variation than real images, and that the variation difference increases with larger frequencies, we introduce a new augmentation scheme, PASTA, that injects variation into the amplitude spectra of synthetic images to help close the syn-to-real gap. PASTA proposes perturbing the amplitude spectra of images in a manner that is proportional to the spatial frequencies, *i.e.*, higher frequencies are perturbed more compared to lower frequencies. If $g_\Lambda(\cdot)$ denotes a perturbation function that returns a perturbed amplitude $\hat{\mathcal{A}}(x)$, i.e., $\hat{\mathcal{A}}(x) = g_\Lambda(\mathcal{A}(x))$, then $g_\Lambda(\cdot)$ for PASTA can be expressed as,

$$g_\Lambda(\mathcal{A}(x))[m,n] = \epsilon[m,n]\mathcal{A}(x)[m,n] \tag{4}$$

$$\text{where } \epsilon[m,n] \sim \mathcal{N}(1, \sigma^2[m,n]) \tag{5}$$

For every spatial frequency $[m,n]$, $\epsilon[m,n]$ ensures a "multiplicative" jitter interaction and is drawn from a gaussian dependent on the spatial frequency. $\sigma[m,n]$ controls the strength of perturbation applied for every spatial frequency. Note that a *naive* uniform perturbation to the amplitude spectrum can be applied with a constant function, $\sigma[m,n] = \beta$ for all $[m,n]$. This results in equal perturbation of all spatial frequencies. To ensure that we perturb HF components more relative to the LF ones, we need to make the variance ($\sigma^2[m,n]$) depend on frequency ($[m,n]$).

For a given frequency we could consider a linear dependence function such as $\sigma[m,n] = 2\alpha\sqrt{\frac{m^2+n^2}{H^2+W^2}} + \beta$, where $2\sqrt{\frac{m^2+n^2}{H^2+W^2}}$ computes the normalized spatial frequency. However, in our empirical observations we found that a linear dependence on frequency does not allow for significant enough growth of perturbation as frequencies increase. Instead we propose the following polynomial function of frequency to allow for sufficient perturbation increases for the high frequency components.

$$\sigma[m,n] = \underbrace{\left(2\alpha\sqrt{\frac{m^2+n^2}{H^2+W^2}}\right)^k}_{\text{proportional to } [m,n]} + \underbrace{\beta}_{\text{uniform}} \tag{6}$$

$\Lambda = \{\alpha, k, \beta\}$ are controllable hyper-parameters. Overall, β ensures a baseline level of jitter applied to all frequencies and α, k govern how the perturbations grow with increasing frequencies. Note that setting either $\alpha = 0$ or $k = 0$ (removing the frequency dependence) results in a setting where the $\sigma[m,n]$ is the same across all $[m,n]$. In appendix, we verify that PASTA augmentation increases the variance metric measured in Fig. 2 [Right] for synthetic images across fine-grained frequency band discretizations.

The steps involved in obtaining a PASTA augmented view are summarized in Alg. 1 and Fig. 3. Given an input image, we first obtain the fourier, amplitude and phase spectra via 2D FFT and then perturb the amplitude spectrum (while ensuring stronger perturbations for HF components) according to Eqns. 4, 5 and 6. Finally, given the perturbed amplitude spectrum and the pristine phase spectrum, we retrieve the augmented image via inverse 2D FFT. In the next section we empricially validate our augmentation strategy.

4. Experimental Details

We conduct synthetic-to-real generalization experiments across three tasks – Semantic Segmentation (SemSeg), Object Detection (ObjDet) and Object Recognition (ObjRec).

[2] In Fig. 2 [Right], for every image, upon obtaining the amplitude spectrum, we first take an element-wise logarithm. Then, for a particular frequency band (pre-defined), we compute the standard deviation of amplitude values within that band (across all the channels). Finally, we average these standard deviations across images to report the same in the bar plots.

[3] When PASTA is applied, we find that performance on "road" increases by a significant margin (per-class generalization results in appendix).

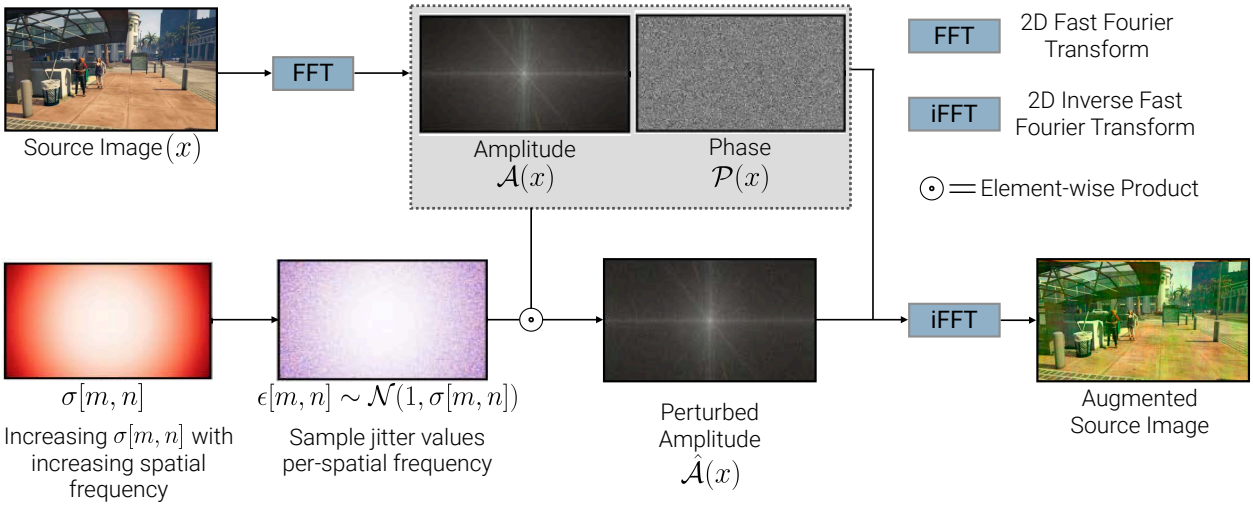

Figure 3: **PASTA.** The figure outlines the augmentation pipeline involved in PASTA. Given an image, we first apply a 2D Fast Fourier Transform (FFT) to obtain the amplitude and phase spectra. Following this, the amplitude spectrum is perturbed as outlined in Eqns. 4, 5 and 6. Finally, we use the perturbed amplitude and the pristine phase spectrum to recover the augmented image by using inverse 2D FFT.

Algorithm 1 PASTA Augmented Views

Input: $x \in \mathbb{R}^{H \times W}$ ▷ Synthetic image (single channel)
PASTA parameters: α, k, β ▷ Set values
$\mathcal{F}(x) \leftarrow \text{FFT}\ (x)$ ▷ Obtain Fourier spectrum
$\mathcal{A}(x) \leftarrow \text{Abs}\ (\mathcal{F}(x))$ ▷ Obtain amplitude spectrum
$\mathcal{P}(x) \leftarrow \text{Ang}\ (\mathcal{F}(x))$ ▷ Obtain phase spectrum
$\mathcal{A}(x) \leftarrow \text{FFTShift}\ (\mathcal{A}(x))$ ▷ Zero-center amplitude
$\sigma_{H \times W} \leftarrow \text{Meshgrid}([-H/2, H/2], [-W/2, W/2])$
for $m \in [-H/2, H/2]$ **do** ▷ Vectorized in practice
  for $n \in [-W/2, W/2]$ **do**
    $\sigma[m,n] \leftarrow \left(2\alpha \sqrt{\frac{m^2+n^2}{H^2+W^2}}\right)^k + \beta$
$\epsilon_{H \times W} \sim \mathcal{N}(1, \sigma^2_{H \times W})$ ▷ Draw perturbations
$\hat{\mathcal{A}}(x) \leftarrow \epsilon \odot \mathcal{A}(x)$ ▷ Multiplicative jitter
$\hat{\mathcal{A}}(x) \leftarrow \text{FFTShift}\ (\hat{\mathcal{A}}(x))$
$\hat{x} \leftarrow \text{Inverse-FFT}\ (\hat{\mathcal{A}}(x), \mathcal{P}(x))$ ▷ Augmented Image

4.1. Datasets and Shifts

Semantic Segmentation. For SemSeg, we use GTAV [58] as our synthetic source dataset with $\sim$ 25k ground-view images and 19 annotated classes, which are compatible with the classes in real target datasets – Cityscapes [18], BDD100K [76] and Mapillary [48]. We train our models on the training split of the synthetic sources, evaluate on the validation splits of the real targets and report performance using mIoU (mean intersection over union). We train SegFormer and HDRA (source-only) on the entirety of GTAV.

Object Detection. For ObjDet, we use Sim10K [33] as our synthetic source dataset with $\sim$ 10k street-view images (from GTAV [58]) and Cityscapes [18] as our (real) target dataset. Following prior work [34], we train on the entirety of Sim10K to detect instances of "car" and report performance on the val split of Cityscapes using mAP@50 (mean average precision at an IoU threshold of 0.5).

Object Recognition. For ObjRec, we use the VisDA-2017 [54] (syn→real) image-classification benchmark with $\sim$ 152k synthetic images (3D renderings of objects) and $\sim$ 55k real images (from COCO [44]) across 12 classes. We use (class-balanced) accuracy as our evaluation metric.

4.2. Models and Baselines

Models. We use DeepLabv3+ [13] (with backbones ResNet-50 [26] and ResNet-101 [26]), SegFormer [69], and HRDA [29] (both with MiT-B5 backbones) architectures for SemSeg experiments. For ObjDet, we use the Faster-RCNN [57] architecture with ResNet-50 and ResNet-101 backbones. For ObjRec, we use ViT-B/16 [20] and ResNet-101 architectures, with both supervised and self-supervised (DINO [6]) initializations. We set PASTA hyper-parameters $(\alpha = 3.0, k = 2.0, \beta = 0.25)$ for SemSeg, ObjDet and ObjRec across shifts, backbones and apply it in conjunction with consistent geometric and photometric augmentations per task. We provide more details in appendix.

Points of Comparison. In addition to prior work in syn-to-real generalization, we also compare PASTA with other augmentation strategies – (1) RandAugment (RandAug) [19] and (2) Photometric Distortion (PD) [11]. The sequence of operations in PD to augment input images are – randomized

Method	Real mIoU ↑				
	G→C	G→B	G→M	Avg	Δ
ResNet-50					
1 Baseline (B) [17]*	28.95	25.14	28.18	27.42	
2 B + RandAug [19]	31.89	38.28	34.54	34.54±0.57	+7.12
3 B + PASTA	**44.12**	**40.19**	**47.11**	**43.81**±0.74	+16.39
ResNet-101					
4 Baseline (B) [17]*	32.97	30.77	30.68	31.47	
5 B + PASTA	**45.33**	**42.32**	**48.60**	**45.42**±0.14	+13.95

Table 1: **PASTA considerably improves a (SemSeg) baseline.** Semantic Segmentation DeepLabv3+ models trained on GTAV (G) and evaluated on {Cityscapes (C), BDD100K (B), Mapillary (M)}. * indicates numbers drawn from published manuscripts. **Bold** indicates best. Δ indicates (absolute) improvement over Baseline. ± indicates the standard deviation across 3 random seeds.

Method	mAP@50 ↑	Δ
ResNet-50	S→C	
1 Baseline	39.4	
2 Baseline + PD [11]	51.5	+12.1
3 Baseline + RandAug [19]	52.8	+13.4
4 Baseline + PASTA	56.3	+16.9
5 Baseline + PD + PASTA	58.0	+18.6
6 Baseline + RandAug + PASTA	**58.3**	+18.9
ResNet-101		
7 Baseline	43.3	
8 Baseline + PD [11]	52.2	+8.9
9 Baseline + RandAug [19]	57.2	+13.9
10 Baseline + PASTA	55.2	+11.9
11 Baseline + PD + PASTA	56.6	+13.3
12 Baseline + RandAug + PASTA	**59.9**	+16.6

Table 2: **PASTA considerably improves a (ObjDet) baseline.** Object Detection Faster-RCNN models trained on Sim10K (S) and evaluated on Cityscapes (C). * indicates numbers drawn from published manuscripts. Highest is **Bolded**. Δ indicates (absolute) improvement over Baseline.

brightness, randomized contrast, RGB→HSV conversion, randomized saturation & hue changes, HSV→RGB conversion, randomized contrast, and randomized channel swap.

5. Results and Findings

5.1. Synthetic-to-Real Generalization Results

Our syn-to-real generalization results for Semantic Segmentation (SemSeg), Object Detection (ObjDet) and Object Recognition (ObjRec) are summarized in Tables. 1, 2, 3 4, 5, 6, and 8. We discuss these results below.

Method	Arch.	Init.	Accuracy ↑ Syn→Real	Δ
1 Baseline	ResNet-101	Sup.	47.22	
2 Baseline + PASTA	ResNet-101	Sup.	**54.39**	+7.17
3 Baseline	ViT-B/16	Sup.	56.06	
4 Baseline + PASTA	ViT-B/16	Sup.	**58.08**	+2.02
5 Baseline	ViT-B/16	DINO [6]	60.93	
6 Baseline + PASTA	ViT-B/16	DINO [6]	**63.55**	+2.62

Table 3: **PASTA considerably improves a (ObjRec) baseline.** Classification models trained on the synthetic source split of VisDA-C and evaluated on the real-split of VisDA-C. **Bold** is best. Δ indicates (absolute) improvement over Baseline. Sup.= Supervised.

▷ **PASTA considerably improves a vanilla baseline.** Tables. 1 and 2 show the improvements offered by PASTA for Semantic Segmentation (SemSeg) and Object Detection (ObjDet) respectively when applied to a vanilla baseline. For SemSeg (see Table. 1), we find that PASTA improves a baseline DeepLabv3+ model by 13+ absolute mIoU points (see rows 1, 3, 4 and 5) across R-50 and R-101 backbones. Furthermore, these improvements are obtained consistently across *all* real target datasets. Similarly, for ObjDet (Table. 2), PASTA offers absolute improvements of 11+ mAP points for a Faster-RCNN baseline across R-50 and R-101 backbones (see rows 1, 4, 7 and 10). We further note that for SemSeg, PASTA outperforms RandAugment [19], a competing augmentation strategy. For ObjDet, PASTA either outperforms (R-50; rows 3, 4) or is competitive with RandAug (R-101; rows 9, 10). For ObjRec (see Table. 3), we find that PASTA significantly improves a vanilla baseline across multiple architectures – for R-101 (rows 1, 2) and ViT-B/16 (rows 3, 4) – and initializations – for supervised (rows 3, 4) and DINO [6] (rows 5, 6) ViT-B/16 initializations.

▷ **PASTA outperforms state-of-the-art generalization methods.** Table. 4 shows how applying PASTA to a baseline outperforms existing generalization methods for SemSeg. Baseline + PASTA outperforms IBN-Net [53], ISW [17], DRPC [77], ASG [15], CSG [14], WEDGE [36], FSDR [31], WildNet [39], DIRL [70] & SHADE [78] in terms of average mIoU across real targets (for both R-50 and R-101). We would like to note that DRPC, ASG, CSG, WEDGE, FSDR, WildNet & SHADE (for R-101) use either more synthetic or real data (the entirety of GTAV [58] or additional datasets) or different base architectures, making these comparisons unfair *to* PASTA. For instance, WEDGE and CSG use DeepLabv2, ASG uses FCNs, DRPC & SHADE (for R-101) use the entirety of GTAV (not just the training split;

Method	Real mIoU ↑				
	G→C	G→B	G→M	Avg	Δ
ResNet-50					
1 IBN-Net [53]*	33.85	32.30	37.75	34.63	
2 ISW [17]*	36.58	35.20	40.33	37.37	
3 DRPC [77]*	37.42	32.14	34.12	34.56	
4 WEDGE [36]*	38.36	37.00	44.82	40.06	
5 ASG [15]*	31.89	N/A	N/A	N/A	
6 CSG [14]*	35.27	N/A	N/A	N/A	
7 WildNet [39]*	44.62	38.42	46.09	43.04	
8 DIRL [70]*	41.04	39.15	41.60	40.60	
9 SHADE [78]*	**44.65**	39.28	43.34	42.42	
10 B + PASTA	44.12	**40.19**	**47.11**	**43.81**±0.74	+0.77
ResNet-101					
11 IBN-Net [53]*	37.37	34.21	36.81	36.13	
12 ISW [17]*	37.20	33.36	35.57	35.58	
13 DRPC [77]*	42.53	38.72	38.05	39.77	
14 WEDGE [36]*	45.18	41.06	48.06	44.77	
15 ASG [15]*	32.79	N/A	N/A	N/A	
16 CSG [14]*	38.88	N/A	N/A	N/A	
17 FSDR [31]*	44.80	41.20	43.40	43.13	
18 WildNet [39]*	45.79	41.73	47.08	44.87	
19 SHADE [78]*	**46.66**	**43.66**	45.50	45.27	
20 B + PASTA	45.33	42.32	**48.60**	**45.42**±0.14	+0.15

Table 4: **PASTA outperforms SOTA (SemSeg) generalization methods.** Semantic Segmentation DeepLabv3+ models trained on GTAV (G) and evaluated on {Cityscapes (C), BDD100K (B), Mapillary (M)}. * indicates numbers drawn from published manuscripts. **Bold** is best. Δ indicates (absolute) improvement over SOTA. ± indicates the standard deviation across 3 random seeds. Rows highlighted in gray use different base architectures and / or extra training data and have been included for completeness. B = Baseline.

2× more data compared to PASTA) and WEDGE uses ∼5k extra Flickr images in it's overall pipeline. FSDR uses FCNs and the the entirety of GTAV for training. FSDR and WildNet also use extra ImageNet [37] images for stylization / randomization. For FSDR, the first step in the pipeline also requires access to SYNTHIA [59]. Unlike PASTA, FSDR and DRPC also select the checkpoints that perform best on target data [1], which is unrealistic since assuming access to labeled (real) target data (for training or model selection) is not practical for syn-to-real generalization. We note that despite having access to *less* data, PASTA outperforms these methods on GTAV→Real. For ObjDet, in Table. 5, we find that combining RandAug + PASTA sets new state-of-the-art on Sim10K [33]→Cityscapes [18].

▷ **PASTA outperforms state-of-the-art adaptation methods.** In Table. 5, we find that both Baseline + PASTA and Baseline + RandAug + PASTA significantly outperform state-of-the-art adaptive object detection method, AWADA [45] (rows 2, 3 and 7). Note that unlike the methods in rows 4-7, PASTA does not have access to real images during training!

▷ **PASTA is complementary to existing generalization methods.** In addition to ensuring that a baseline model

Method	Real Data	mAP@50 ↑	Δ
Generalization		S→C	
1 Baseline (B)	✗	43.3	
2 B + PASTA	✗	55.2	+11.9
3 B + RandAug + PASTA	✗	**59.9**	+16.6
(Unsupervised) Adaptation			
4 EPM [30]*	✓	51.2	+7.9
5 Faster-RCNN w/ rot [67]*	✓	52.4	+9.1
6 ILLUME [34]*	✓	53.1	+9.8
7 AWADA [45]*	✓	53.2	+9.9
8 Faster-RCNN (Oracle) [67]*	✓	70.4	+27.1

Table 5: **PASTA outperforms SOTA (ObjDet) adaptation methods.** Object Detection Faster-RCNN (ResNet-101) models trained on Sim10K (S) and evaluated on Cityscapes (C). * indicates numbers drawn from published manuscripts. **Bold** is best. Rows in gray are adaptation methods with access to unlabeled real data. Rows in red have access to labeled real data. Δ indicates (absolute) improvement over Baseline (B). PASTA leads to outperforming adaptation methods without any access to real data during training.

Method	Real mIoU ↑				
	G→C	G→B	G→M	Avg	Δ
ResNet-50					
1 IBN-Net [53]*	33.85	32.30	37.75	34.63	
2 IBN-Net + PASTA	**41.90**	**41.46**	**45.88**	**43.08**±0.37	+8.45
3 ISW [17]*	36.58	35.20	40.33	37.37	
4 ISW + PASTA	**42.13**	**40.95**	**45.67**	**42.91**±0.27	+5.54
ResNet-101					
5 IBN-Net [53]*	37.37	34.21	36.81	36.13	
6 IBN-Net + PASTA	**43.64**	**42.46**	**47.51**	**44.54**±0.89	+8.41
7 ISW [17]*	37.20	33.36	35.57	35.58	
8 ISW + PASTA	**44.46**	**43.02**	**47.35**	**44.95**±0.21	+9.37

Table 6: **PASTA is complementary to existing (SemSeg) generalization methods.** Semantic Segmentation generalization methods (DeepLabv3+ models) trained on GTAV (G) and evaluated on {Cityscapes (C), BDD100K (B), Mapillary (M)}. * indicates numbers drawn from published manuscripts. **Bold** is best. Δ indicates (absolute) improvement over the base generalization method. ± indicates the standard deviation across 3 random seeds.

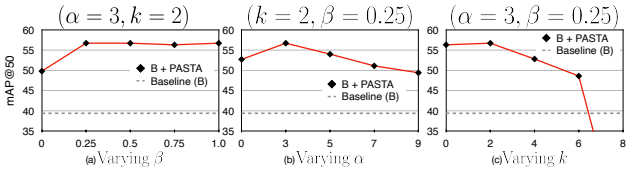

Figure 4: **Sensitivity of α, β, k in PASTA for Object Detection.** Faster-RCNN (ResNet-50) models trained on Sim10K (S) and evaluated on Cityscapes (C). We vary α, k and β for PASTA within the sets $\{0, 3, 5, 7, 9\}$, $\{0, 2, 4, 6, 8\}$ and $\{0, 0.25, 0.5, 0.75, 1.0\}$.

improves over existing methods, we find that PASTA is also complementary to existing generalization methods. For SemSeg, in Table. 6, we find that applying PASTA significantly improves generalization performance (5+ absolute mIoU points) of IBN-Net [53] and ISW [17] across R-50 and R-101. For ObjDet, we find that PASTA is complementary to existing augmentation methods (PD [11] and RandAug [19]; rows 5, 6, 11 and 12 in Table. 2), with RandAug + PASTA setting state-of-the-art on Sim10K [33]→Cityscapes [18]. In appendix, we applied PASTA to CSG [14], a state-of-the-art generalization method for ObjRec on VisDA-C [54] CSG already utilizes RandAugment, so we tested PASTA in two settings: with and without RandAugment. In both scenarios, incorporating PASTA led to improved performance.

5.2. Analyzing PASTA

▷ **Sensitivity of PASTA to α, k and β.** While β provides a baseline level of uniform jitter in the frequency domain, α, k govern the degree of monotonicity applied to the perturbations. To assess the sensitivity of PASTA to α, k, β, in Fig. 4, we conduct experiments for ObjDet where we vary one hyper-parameter while freezing the other two. We find that performance is stable when β exceeds a certain threshold. For α, we find that while performance drops with increasing α, worst generalization performance is still significantly above baseline. More importantly, we find generalization improvements offered by PASTA are sensitive to "extreme" values of k. Qualitatively, overly high values of α, k lead to augmented views which have their semantic content significantly occluded, thereby resulting in poor generalization. As an anecdotal rule of thumb, for a vanilla baseline that is not designed specifically for syn-to-real generalization, we find that restricting $k \in [1, 4]$ leads to stable improvements.

▷ **PASTA vs other frequency-based augmentation strategies.** Prior work has also considered augmenting images in the Fourier domain for syn-to-real generalization (FSDR [31]), multi-source domain generalization (FACT [71]), domain adaptation (FDA [73]) and robustness (APR [10]). Row 17 in Table. 4 already shows how PASTA outperforms FSDR. In Table. 7, we conduct a control

Method	Real Data	Real mIoU	Δ
1 Baseline	✗	26.99	
2 Baseline + FDA [73]	✓	33.04	+6.05
3 Baseline + APR-P [10]	✗	37.52	+10.53
4 Baseline + AJ (FACT [71])	✗	30.70	+3.71
5 Baseline + AM (FACT [71])	✗	39.70	+12.71
6 Baseline + PASTA	✗	**41.90**	+14.91

Table 7: **PASTA vs Frequency-domain Augmentations.** Semantic Segmentation DeepLabv3+ (R-50) models trained with different frequency domain augmentation strategies on GTAV (at an input resolution of 1024×560 due to compute constraints) and evaluated on {Cityscapes, BDD100K, Mapillary}. **Bold** indicates best. Δ indicates (absolute) improvement over Baseline. Row in gray uses real data for augmenting images.

experiment to compare PASTA with some other frequency domain augmentation strategies (summarized below) when applied to a baseline DeepLabv3+ (R-50) SemSeg model for the GTAV→Real shift (we downsample input images to a resolution of 1024×560 during training for all methods due to limited computational resources).

- **FDA [73] –** FDA is a recent approach for syn-to-real domain adaptation and naturally requires access to unlabeled target data. In FDA, to generate augmented views, low frequency bands of the amplitude spectra of source images are replaced with those of target – essentially mimicking a cheap style transfer operation. Since we do not assume access to target data in our experimental settings, a direct comparison is not possible. Instead, we consider a proxy task where we intend to generalize to real datasets (Cityscapes, BDD100K, Mapillary) by assuming additional access to 6 real world street view images under different weather conditions (for style transfer) – sunny day, rainy day, cloudy day, etc. – in addition to synthetic images from GTAV. We find that PASTA outperforms FDA (row 2 vs row 6 in Table. 7).
- **APR-P [10] –** Amplitude Phase Recombination is a recent method designed to improve robustness against natural corruptions. APR replaces the amplitude spectrum of an image with the amplitude spectrum from an augmented view (APR-S) or different images (APR-P). When applied to synthetic images from GTAV, we find that PASTA outperforms APR-P (row 3 vs row 6 in Table. 7).
- **FACT [71] –** FACT is a multi-source domain generalization method for object recognition that uses one of two frequency domain augmentation strategies – Amplitude Jitter (AJ) and Amplitude Mixup (AM) – in a broader training pipeline. AM involves perturbing the amplitude spectrum of the image of concern by performing a convex combination with the amplitude spectrum of another "mixup"

Method	G→C mIoU ↑	Δ
DeepLabv3+ (ResNet-101)		
1 Baseline	31.47	
2 Baseline + PASTA	**45.42**	+13.95
SegFormer (MiT B5)		
3 Baseline [28]	45.60	
4 Baseline + PASTA	**52.57**	+6.97
HRDA (MiT B5)		
5 Baseline [29]	53.01	
6 Baseline + PASTA	**57.21**	+4.20

Table 8: **PASTA is agnostic to choice of (SemSeg) architecture.** Semantic Segmentation models trained on GTAV (G) and evaluated on Cityscapes (C). **Bold** indicates best. Δ indicates (absolute) improvement over Baseline.

image drawn from the same source data. AJ (AJ) perturbs the amplitude spectrum with a single jitter value ϵ for all spatial frequencies and channels. We compare with both AM and AJ for SemSeg and find that PASTA outperforms both (rows 4, 5 vs row 6 in Table. 7). Additionally, for multi-source domain generalization on PACS [40], we find that FACT-PASTA (where we replace the augmentations in FACT with PASTA) outperforms FACT-Vanilla – 87.97% vs 87.10% average leave-one-out domain accuracy.

▷ **Does monotonicity matter in PASTA?** As stated earlier, a key insight while designing PASTA was to perturb the high-frequency components in the amplitude spectrum more relative to the low-frequency ones. This monotonicity is governed by the choice of α, k in Eqn. 6 (β applies a uniform level of jitter to all frequency components). To assess the importance of this monotonic setup, we compare generalization improvements offered by PASTA by comparing $\alpha = 0$ (uniform) and $\alpha = 3$ (monotonically increasing) settings. For SemSeg, for a baseline DeepLabv3+ (R-50) model, we find that $\alpha = 3$ improves over $\alpha = 0$ by 3.89 absolute mIoU points (see Table. 1 for experimental setting). Similarly, for ObjDet, for a baseline Faster-RCNN (R-50) model, we find that $\alpha = 3$ improves of $\alpha = 0$ by 3.4 absolute mAP points (see Table. 2 for experimental setting). Additionally, reversing the monotonic trend (LF perturbed more than HF) leads to significantly worse generalization performance for SemSeg – 8.90 Avg. Real mIoU, dropping even below vanilla baseline performance of 27.42 mIoU.

▷ **Does PASTA help for transformer based architectures?** Our key syn-to-real generalization results across SemSeg and ObjDet are with CNN backbones. Following the recent surge of interest in introducing transformers to vision tasks [20], we also conduct experiments to assess if syn-to-real generalization improvements offered by PASTA generalize beyond CNNs. In Table. 3, we show how applying PASTA improves syn-to-real generalization performance of ViT-B/16 baselines for both supervised (rows 3, 4) and self-supervised DINO [6] initializations (rows 5, 6). In Table. 8, we consider SegFormer [69] and HRDA [29] (source-only), recent transformer based segmentation frameworks, and check syn-to-real generalization performance when trained on GTAV [58] and evaluated on Cityscapes [18]. We find that applying PASTA significantly improves performance of a vanilla baseline (6+ and 4+ absolute mIoU points for SegFormer and HRDA respectively).

To summarize, from our experiments, we find that PASTA serves as a simple and effective *plug-and-play* augmentation strategy for training on synthetic data that

- **provides strong out-of-the-box generalization performance** – enables a baseline method to outperform existing generalization approaches
- **is complementary to existing generalization methods** – applying PASTA to existing methods or augmentation strategies leads to improvements
- **is applicable across tasks, backbones and shifts** – PASTA leads to improvements across SemSeg, ObjDet, ObjRec for multiple backbones for five syn-to-real shifts

6. Conclusion

We propose Proportional Amplitude Spectrum Training Augmentation (PASTA), a *plug-and-play* augmentation strategy for synthetic-to-real generalization. PASTA is motivated by the observation that the amplitude spectra are less diverse in synthetic than real data, especially for high-frequency components. Thus, PASTA augments synthetic data by perturbing the amplitude spectra, with magnitudes increasing for higher frequencies. We show that PASTA offers strong out-of-the-box generalization performance on semantic segmentation, object detection, and object classification tasks. The strong performance of PASTA holds true alone (i.e., training with ERM using PASTA augmented images) or together with alternative generalization/augmentation algorithms. We would like to emphasize that the strength of PASTA lies in its simplicity (just modify your augmentation pipeline) and effectiveness, offering strong improvements despite not using extra modeling components, objectives, or data. We hope that future research endeavors in syn-to-real generalization take augmentation techniques like PASTA into account. Additionally, it might be of interest to the research community to explore how PASTA could be utilized for *adaptation* – when "little" real target is available.

Acknowledgments. We thank Viraj Prabhu and George Stoica for fruitful discussions and valuable feedback. This work was supported in part by sponsorship from NSF Award #2144194, NASA ULI, ARL, and Google.

References

[1] Largely lower results on bdd and maphillary · issue #2 · jxhuang0508/fsdr. https://github.com/jxhuang0508/FSDR/issues/2. (FSDR Official Code Repository). 7

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 2

[3] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. In *Advances in Neural Information Processing Systems*, pages 998–1008, 2018. 2

[4] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24, 2011. 2

[5] Geoffrey J Burton and Ian R Moorhead. Color and spatial structure in natural scenes. *Applied optics*, 26(1):157–170, 1987. 3

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 5, 6, 9

[7] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[8] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *European Conference on Computer Vision*, pages 301–318. Springer, 2020. 2

[9] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8866–8875, 2020. 2

[10] Guangyao Chen, Peixi Peng, Li Ma, Jia Li, Lin Du, and Yonghong Tian. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 458–467, 2021. 3, 8

[11] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5, 6, 8

[12] Keyu Chen, Di Zhuang, and J Morris Chang. Discriminative adversarial domain generalization with meta-learning based cross-domain validation. *Neurocomputing*, 467:418–426, 2022. 2

[13] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1, 5

[14] Wuyang Chen, Zhiding Yu, Shalini De Mello, Sifei Liu, Jose M. Alvarez, Zhangyang Wang, and Anima Anandkumar. Contrastive syn-to-real generalization. In *International Conference on Learning Representations*, 2021. 2, 6, 7, 8

[15] Wuyang Chen, Zhiding Yu, Zhangyang Wang, and Animashree Anandkumar. Automated synthetic-to-real generalization. In *International Conference on Machine Learning*, pages 1746–1756. PMLR, 2020. 2, 6, 7

[16] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 2

[17] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11580–11590, 2021. 2, 6, 7, 8

[18] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1, 2, 4, 5, 7, 8, 9

[19] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 1, 5, 6, 8

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 5, 9

[21] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *Advances in Neural Information Processing Systems*, pages 6447–6458, 2019. 2

[22] Xinjie Fan, Qifei Wang, Junjie Ke, Feng Yang, Boqing Gong, and Mingyuan Zhou. Adversarially adaptive normalization for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8208–8217, 2021. 2

[23] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559, 2015. 2

[24] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. 2

[25] Bruce C Hansen and Robert F Hess. Structural sparseness and spatial phase alignment in natural scenes. *JOSA A*, 24(7):1873–1885, 2007. 1, 2, 3

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[27] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019. 3

[28] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9924–9935, 2022. 9

[29] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 372–391. Springer, 2022. 5, 9

[30] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *European Conference on Computer Vision*, pages 733–748. Springer, 2020. 7

[31] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6891–6902, 2021. 1, 2, 3, 6, 7, 8

[32] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Rda: Robust domain adaptation via fourier adversarial attacking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8988–8999, 2021. 3

[33] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, page 746–753. IEEE Press, 2017. 2, 5, 7, 8

[34] Vaishnavi Khindkar, Chetan Arora, Vineeth N Balasubramanian, Anbumani Subramanian, Rohit Saluja, and CV Jawahar. To miss-attend is to misalign! residual self-attentive feature alignment for adapting object detectors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3632–3642, 2022. 5, 7

[35] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012. 2

[36] Namyup Kim, Taeyoung Son, Jaehyun Pahk, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Wedge: Web-image assisted domain generalization for semantic segmentation. In *2023 International Conference on Robotics and Automation (ICRA)*, 2023. 2, 6, 7

[37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 7

[38] Jogendra Nath Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, and R Venkatesh Babu. Generalize then adapt: Source-free domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7046–7056, 2021. 2

[39] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. Wildnet: Learning domain generalized semantic segmentation from the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9936–9946, 2022. 6, 7

[40] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5542–5550, 2017. 9

[41] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[42] Lei Li, Ke Gao, Juan Cao, Ziyao Huang, Yepeng Weng, Xiaoyue Mi, Zhengze Yu, Xiaoya Li, and Boyang Xia. Progressive domain expansion network for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 224–233, 2021. 2

[43] Xiaotong Li, Yongxing Dai, Yixiao Ge, Jun Liu, Ying Shan, and Ling-Yu Duan. Uncertainty modeling for out-of-distribution generalization. In *International Conference on Learning Representations*, 2022. 2

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5

[45] Maximilian Menke, Thomas Wenzel, and Andreas Schwung. Awada: Attention-weighted adversarial domain adaptation for object detection, 2022. 7

[46] S. Mishra, R. Panda, C. Phoo, C. Chen, L. Karlinsky, K. Saenko, V. Saligrama, and R. S. Feris. Task2sim: Towards effective pre-training and transfer from synthetic data. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9184–9194, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society. 2

[47] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013. 2

[48] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. 2, 4, 5

[49] Oren Nuriel, Sagie Benaim, and Lior Wolf. Permuted adain: Reducing the bias towards global statistics in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[50] Henri J Nussbaumer. The fast fourier transform. In *Fast Fourier Transform and Convolution Algorithms*, pages 80–111. Springer, 1981. 3

[51] A.V. Oppenheim and J.S. Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981. 1, 2, 3

[52] A. Oppenheim, Jae Lim, G. Kopec, and S. Pohlig. Phase in speech and pictures. In *ICASSP '79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 632–637, 1979. 1, 2, 3

[53] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 2, 6, 7, 8

[54] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017. 2, 4, 5, 8

[55] Leon N Piotrowski and Fergus W Campbell. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception*, 11(3):337–346, 1982. 1, 2, 3

[56] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020. 2

[57] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 5

[58] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016. 1, 2, 3, 4, 5, 6, 9

[59] 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. 7

[60] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. In *International Conference on Learning Representations*, 2018. 2

[61] Swami Sankaranarayanan, Yogesh Balaji, Carlos D. Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 4

[62] Yuge Shi, Jeffrey Seely, Philip Torr, Siddharth N, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. In *International Conference on Learning Representations*, 2022. 2

[63] DJ Tolhurst, Yv Tadmor, and Tang Chao. Amplitude spectra of natural images. *Ophthalmic and Physiological Optics*, 12(2):229–232, 1992. 3

[64] V. S. Vibashan, Vikram Gupta, Poojan Oza, Vishwanath A. Sindagi, and Vishal M. Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4514–4524, 2021. 2

[65] Bailin Wang, Mirella Lapata, and Ivan Titov. Meta-learning for domain generalization in semantic parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–379, Online, June 2021. Association for Computational Linguistics. 2

[66] Jingye Wang, Ruoyi Du, Dongliang Chang, Kongming Liang, and Zhanyu Ma. Domain generalization via frequency-domain-based feature disentanglement and interaction. MM '22, page 4821–4829, New York, NY, USA, 2022. Association for Computing Machinery. 3

[67] Xin Wang, Thomas E Huang, Benlin Liu, Fisher Yu, Xiaolong Wang, Joseph E Gonzalez, and Trevor Darrell. Robust object detection via instance-level temporal cycle confusion. *International Conference on Computer Vision (ICCV)*, 2021. 7

[68] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 834–843, 2021. 2

[69] 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, 9

[70] Qi Xu, Liang Yao, Zhengkai Jiang, Guannan Jiang, Wenqing Chu, Wenhui Han, Wei Zhang, Chengjie Wang, and Ying Tai. Dirl: Domain-invariant representation learning for generalizable semantic segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2884–2892, Jun. 2022. 6, 7

[71] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021. 1, 2, 3, 4, 8

[72] Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. Phase consistent ecological domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9011–9020, 2020. 1

[73] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 1, 3, 8

[74] Yijun Yang, Shujun Wang, Pheng-Ann Heng, and Lequan Yu. Hcdg: A hierarchical consistency framework for domain generalization on medical image segmentation. *arXiv preprint arXiv:2109.05742*, 2021. 2

[75] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[76] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 2, 4, 5

[77] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2100–2110, 2019. 2, 6, 7

[78] Yuyang Zhao, Zhun Zhong, Na Zhao, Nicu Sebe, and Gim Hee Lee. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 6, 7

[79] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2021. 2