Generalizable Fourier Augmentation for Unsupervised Video Object Segmentation

Huihui Song[1*], Tiankang Su[1*†], Yuhui Zheng[1,2], Kaihua Zhang[1], Bo Liu[3], Dong Liu[4]

[1]B-DAT and CICAEET, Nanjing University of Information Science and Technology, Nanjing, China
[2] College of Computer, Qinghai Normal University, Xining 810016, China
[3] Walmart Global Tech, Sunnyvale, CA, 94086, USA
[4] Netflix Inc, Los Gatos, CA, 95032, USA
tiankangsu@gmail.com

Abstract

The performance of existing unsupervised video object segmentation methods typically suffers from severe performance degradation on test videos when tested in out-of-distribution scenarios. The primary reason is that the test data in real-world may not follow the independent and identically distribution (*i.i.d.*) assumption, leading to domain shift. In this paper, we propose a Generalizable Fourier Augmentation (GFA) method during training to improve the generalization ability of the model. To achieve this, the GFA performs Fast Fourier Transform (FFT) over the intermediate spatial domain features in each layer to yield corresponding frequency representations, including amplitude components (encoding scene-aware styles such as texture, color, contrast of the scene) and phase components (encoding rich semantics). We produce a variety of style features via Gaussian sampling to augment the training data, thereby improving the generalization capability of the model. To further improve the cross-domain generalization performance of the model, we design a phase feature update strategy via exponential moving average using phase features from past frames in an online update manner, which could help the model to learn cross-domain-invariant features. Extensive experiments show that the proposed GFA achieves the state-of-the-art performance on popular benchmarks.

Introduction

Given a video sequence, unsupervised video object segmentation (UVOS) (Zhou et al. 2020; Ji et al. 2021; Tokmakov, Alahari, and Schmid 2017; Zhang et al. 2021) aims to locate and segment the primarily moving foreground targets without any prior knowledge. This task setting is the same as zero-shot learning in which the test instances may not be seen during training (Chen et al. 2020). UVOS has been widely applied in a variety of practical applications such as visual tracking , autonomous driving and video surveillance.

The existing UVOS methods (Zhang et al. 2021; Ren et al. 2021) are trained and tested by the samples with an implicit assumption that they are independently and identically distributed (*i.i.d.*), i.e., both training and test data are in-distribution samples. Despite the demonstrated success, in practical applications, the UVOS models (Chen et al. 2020; Yue et al. 2021) often suffer from the out-of-distribution (OOD) test samples due to the zero-shot task setting, which

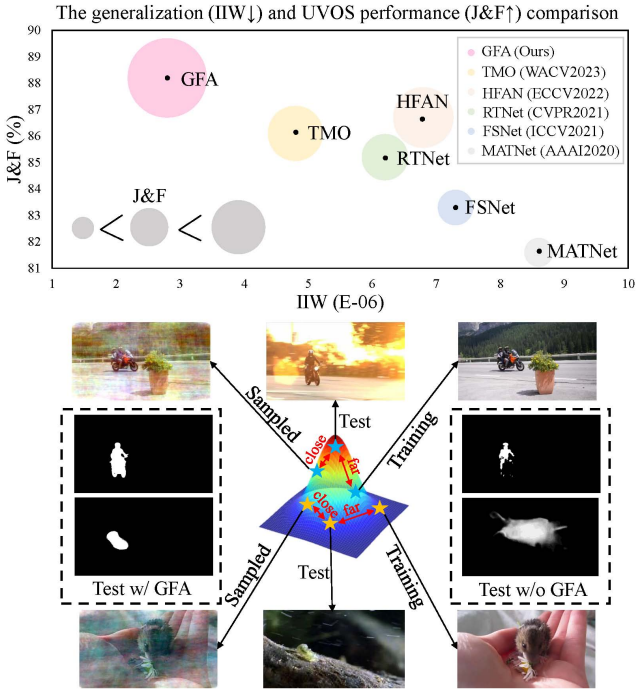

Figure 1: Bottom: Training samples include *motorcycle* and *mouse* categories while the seen test sample *motorcycle* with different scene styles suffers from "*scene shift*" and the unseen test sample *worm* undergoes "*semantic shift*". To reduce both types of distribution shifts, our Generalizable Fourier Augmentation (GFA) augments the features in the frequency domain by sampling the amplitudes from a Gaussian distribution while online updating the phase components across different domains, which helps the model to learn style-rich and cross-domain-invariant features that are effective to improve model generalization capability (i.e., Top: Information stored In Weights (IIW) (Wang et al. 2022)) vs. accuracy measured by $\mathcal{J}\&\mathcal{F}$ commonly used in UVOS between our GFA and the state-of-the-art methods).

*These authors contributed equally.
†Corresponding author.

degrades the model performance significantly. In UVOS, we observe that there are two OOD scenarios that may cause distribution shifts in the test samples. Among them, we term the first scenario as *"scene shift"*, e.g., the same semantic categories of targets in the test videos are seen in the training data, but the targets in the test videos suffer from severe appearance variations due to the varying scene styles of texture, color, and contrast (e.g., the *motorcycle* shown in Figure 1). Besides, we term the second scenario as *"semantic shift"*, where the targets in the test videos might belong to the categories that are not present in the training data (e.g., the novel class of *worm* shown in Figure 1). Without considering these OOD challenges, the existing state-of-the-art UVOS methods (Pei et al. 2022; Cho et al. 2023; Ren et al. 2021; Ji et al. 2021; Zhou et al. 2020) cannot achieve a satisfying generalization capability (see the IIW (Wang et al. 2022) vs. $\mathcal{J}\&\mathcal{F}$ shown in the top of Figure 1).

An intuitive way to improve the generalization capability of a model is to allow it to see diverse samples from different domains (Tobin et al. 2017). Based on this viewpoint, we address the *"scene shift"* issue by making the model see the diverse scene styles in training. Recently, numerous evidences show that the amplitude spectrum of Fourier Transform of a scene image encodes the rich scene style information (e.g., the texture, color, contrast information in the scene) (Oliva and Torralba 2001; Xu et al. 2021; Yang and Soatto 2020). Therefore, properly manipulating the amplitude spectrum of the image Fourier Transform can generate a variety of diverse scene styles that facilitate to improve the model generalization capability. Inspired by this, to generate the diverse scene styles, we propose to augment the amplitude spectrum features with uncertainty via Gaussian sampling. Specifically, given the spatial domain intermediate features of an image, we leverage the Fast Fourior Transform (FFT) (Brigham and Morrow 1967) to decompose the features into amplitude components and phase components. Then, to increase the diversity of styles, we introduce uncertainty modeling with an assumption that the amplitude feature statistics (i.e., mean and standard deviation) follow a Gaussian distribution. In this way, we can produce diverse scene style features to reduce the dependency of the model on the training samples and effectively reduce *"scene shift"*.

Different from the amplitude spectrum that can capture rich scene style information, the phase spectrum of the image Fourier Transform encodes rich semantic information (Oliva and Torralba 2001; Xu et al. 2021; Yang and Soatto 2020), which can be used to address the *"semantic shift"* issue. To achieve this, we propose an online update strategy to process the phase spectrum features. Given the current phase features and the previous phase features from the past frames updated before, we apply an Exponential Moving Average (EMA) technique (Klinker 2011) to update the phase features by the phase features of the past frames. In this way, the smoothed phase features can be correlated with the entire dataset in an online update manner to produce cross-domain-invariant features, which helps reduce the *"semantic shift"*, further improving the generalization capability of the model.

In summary, our main contributions include:

(1) a novel generalizable Fourier augmentation framework to learn robust domain-invariant features, which effectively improves the generalization capability of the model for UVOS.

(2) a sampling method with uncertainty modeling on the amplitude feature, which can produce diverse scene-aware style features to address *"scene shift"* issue.

(3) an online update strategy on the phase features to better learn cross-domain-invariant features, which can alleviate the *"semantic shift"* issue and further enhance the generalization ability of the model.

Methodology

Overview

Figure 2 illustrates the network architecture of our model. The model has a one-stream end-to-end Transformer architecture, which is composed of an encoder with a set of specifically-designed Transformer layers and a decoder composed of convolutional layers and an upscale layer.

At the encoder stage, we first divide the input of the concatenating RGB image and optical flow map into a set of tokens with size 4×4 pixels. Then, the tokens are fed into the overlap patch embedding layer to produce the embeddings. Afterward, the embeddings are fed into 4 Transformer layers with each containing N-layer transformer blocks, each of which consists of an Efficient self-attention module and a Mix-FFN module (Xie et al. 2021), to make the token features capture long-range dependency information. After this, the token features are fed into an overlapping patch merging layer to reconstruct the whole feature map $\mathcal{X} \in \mathbb{R}^{H\times W\times C}$. The feature map $\mathcal{X}$ is transformed by a 2-D FFT to produce the corresponding frequency-domain feature representations of amplitude $\mathcal{A} \in \mathbb{R}^{H\times W\times C}$ and phase $\mathcal{P} \in \mathbb{R}^{H\times W\times C}$. We then sample the amplitude features $\mathcal{A}$ to yield diverse scene-aware features, which can effectively augment the training samples. Besides, we employ an EMA strategy to online update the phase features, which can help to learn cross-domain-invariant features that can enhance the model generalization capability. In the end, the sampled amplitude and online updated phase features are combined and fed into an iFFT layer to reconstruct the augmented spatial features.

At the decoder stage, the augmented spatial features in each Transformer layer are all fed into an upscale layer which consists of a 1×1 convolution layer and an interpolation layer, to produce the features with the same resolution. At the head of the network, a 1×1 convolution layer followed by an upscaling and a sigmoid layer to yield the predicted segmentation mask $\mathbf{P} \in [0, 1]^{H\times W}$.

Generalizable Fourier Augmentation

Figure 3 illustrates the schematic diagram of our GFA. The GFA first transforms the spatial features into frequency domain via FFT, and then augments the amplitude features via Gaussian sampling technique that takes modeling uncertainty into account, and enhances the phase features via EMA that is an online update strategy, which can effectively learn domain-invariant foreground features.

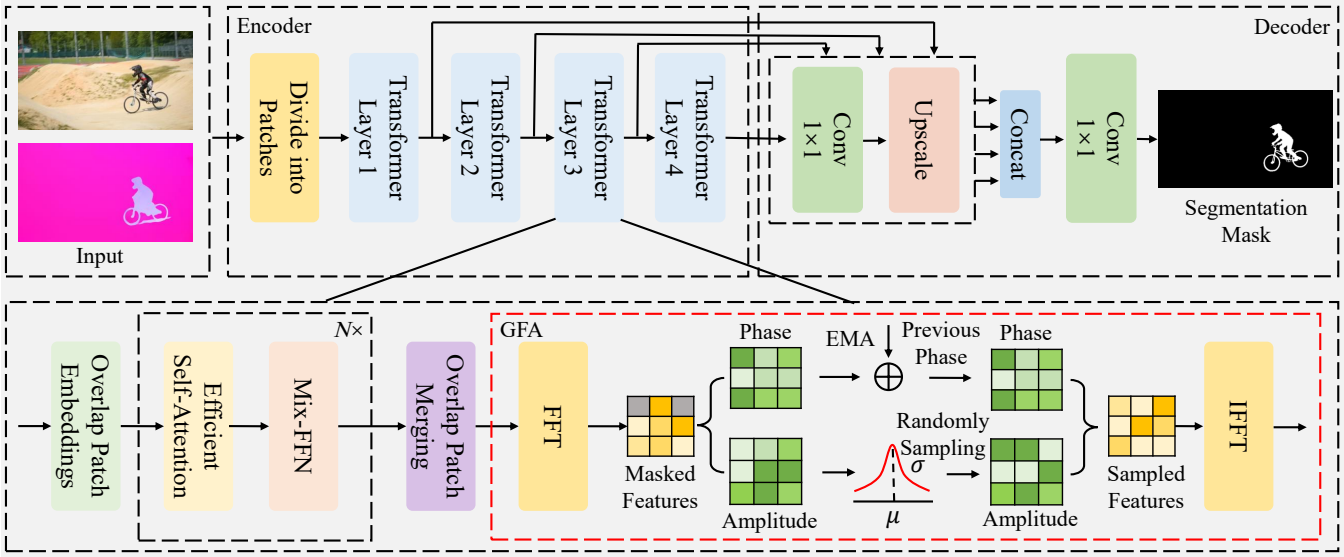

Figure 2: Pipeline of the proposed network architecture for UVOS.

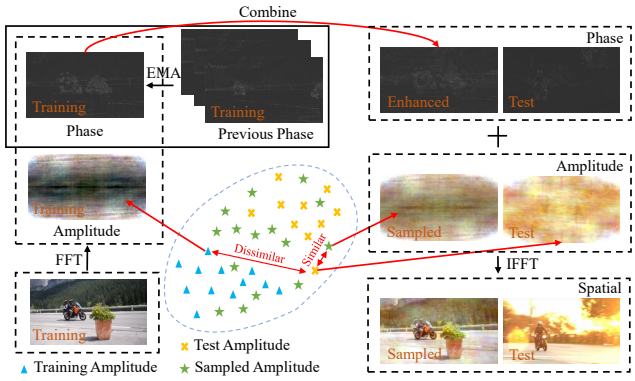

Figure 3: Schematic diagram of the proposed GFA, which includes an amplitude feature sampling process and an online updated phase feature process. The former generates augmented features with diverse scene-aware styles while the latter generates continuously updated domain-invariant foreground features, which are able to enhance the model generalization capability.

Frequency Representation Given the feature map $\mathcal{X} \in \mathbb{R}^{H_i \times W_i \times C_i}$ with width W_i, height H_i, and channel number C_i after the overlap patch merging layer in the i-th Transformer layer, each channel of $\mathcal{X}$ is transformed by FFT to achieve the corresponding frequency representation $\widehat{\mathcal{X}}_i$ as

$$\widehat{\mathcal{X}}_i(x, y) = \mathcal{F}(\mathcal{X}_i), \tag{1}$$

where $\mathcal{F}$ denotes the FFT operator (Brigham and Morrow 1967). Then, $\widehat{\mathcal{X}}_i$ can be decomposed into an amplitude component $\mathcal{A}_i$ and a phase component $\mathcal{P}_i$ as

$$\begin{aligned} \mathcal{A}_i &= \sqrt{R^2(\widehat{\mathcal{X}}_i)(x, y) + I^2(\widehat{\mathcal{X}}_i)(x, y)}, \\ \mathcal{P}_i &= \arctan(I(\widehat{\mathcal{X}}_i)(x, y) / R(\widehat{\mathcal{X}}_i)(x, y)), \end{aligned} \tag{2}$$

where $R(\widehat{\mathcal{X}}_i)$ and $I(\widehat{\mathcal{X}}_i)$ denote the real component and imaginary component of the frequency feature representation $\widehat{\mathcal{X}}_i$, respectively. In addition, combined amplitude $\mathcal{A}_i$ and phase $\mathcal{P}_i$ can be transferred to spatial domain features $\mathcal{X}_i$ via IFFT, which is defined as

$$\mathcal{X}_i = \mathcal{F}^{-1}(\mathcal{A}_i e^{j\mathcal{P}_i}), \tag{3}$$

where $\mathcal{F}^{-1}$ denotes the IFFT operator (Brigham and Morrow 1967).

Amplitude Feature Augmentation The amplitude features $\mathcal{A}_i$ encode rich scene-aware style information about the texture, color, and contrast of the scene (Huang et al. 2021; Oliva and Torralba 2001). An intuitive way to improve the generalization capability of the model to different scenes is to allow it to see diverse scene styles. To this end, we augment the scene-aware samples by manipulating the amplitude statistics (mean and standard deviation). We randomly sample the amplitude statistics from a Gaussian distribution to produce a variety of samples with different styles to reduce the dependency of the model on the training data. As shown in Figure 3, given a spatial domain feature $\mathcal{X}_i$ in the i-th Transformer layer, we decompose $\mathcal{X}_i$ into amplitude components $\mathcal{A}_i$ and phase components $\mathcal{P}_i$ via (2), which denotes the style content and semantics, respectively. Then, we achieve the channel-wise amplitude statistics (mean and standard deviation) in a mini-batch, which is ex-

pressed as

$$\mu(\mathcal{A}_i) = \frac{1}{H_i W_i} \sum_{x=1}^{H_i} \sum_{y=1}^{W_i} \mathcal{A}_i,$$
$$\sigma(\mathcal{A}_i) = \sqrt{\frac{1}{H_i W_i} \sum_{x=1}^{H_i} \sum_{y=1}^{W_i} (\mathcal{A}_i - \mu(\mathcal{A}_i))^2}, \tag{4}$$

where $\mu(\mathcal{A}_i)$ and $\sigma(\mathcal{A}_i)$ are the mean and standard deviation of amplitude $\mathcal{A}_i$, respectively.

The variance of the probability feature statistics models their uncertainty scope (Li et al. 2022; Jeon et al. 2021), and as the variance increases, the amplitude features with the OOD scene styles are more likely to be sampled from the distributions. Inspired by this, we employ a non-parametric estimation method in a mini-batch to obtain the variances of $\mu(\mathcal{A}_i)$ and $\sigma(\mathcal{A}_i)$ for sampling the amplitude features, which can be formulated as

$$\sigma^2_\mu(\mathcal{A}_i) = \frac{1}{B} \sum_{b=1}^{B} (\mu(\mathcal{A}_i) - \mathbb{E}_b(\mu(\mathcal{A}_i)))^2,$$
$$\sigma^2_\sigma(\mathcal{A}_i) = \frac{1}{B} \sum_{b=1}^{B} (\sigma(\mathcal{A}_i) - \mathbb{E}_b(\sigma(\mathcal{A}_i)))^2, \tag{5}$$

where B denotes batch size and $\mathbb{E}_b$ refers to expectation in a mini-batch. $\sigma^2_\mu(\mathcal{A}_i)$, $\sigma^2_\sigma(\mathcal{A}_i)$ represent the variances of $\mu(\mathcal{A}_i)$ and $\sigma(\mathcal{A}_i)$, respectively. Once we have obtained the new feature statistics, i.e. mean $\beta(\mathcal{A}_i) \sim \mathcal{N}(\mu(\mathcal{A}_i), \sigma^2_\mu(\mathcal{A}_i))$ and standard deviation $\gamma(\mathcal{A}_i) \sim \mathcal{N}(\sigma(\mathcal{A}_i), \sigma^2_\sigma(\mathcal{A}_i))$, we can randomly sample rich scene-aware style features from the new distribution $\mathcal{N}(\beta(\mathcal{A}_i), \gamma(\mathcal{A}_i))$ to reduce the model overfitting to the training set. However, the direct sampling operation is not differentiable, and we therefore apply a re-parameterization technique (Kingma and Welling 2013) to make it differentiable as

$$\beta(\mathcal{A}_i) = \mu(\mathcal{A}_i) + \epsilon_\mu \sigma_\mu(\mathcal{A}_i),$$
$$\gamma(\mathcal{A}_i) = \sigma(\mathcal{A}_i) + \epsilon_\sigma \sigma_\sigma(\mathcal{A}_i), \tag{6}$$

where $\epsilon_\mu \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ and $\epsilon_\sigma \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ both follow the standard normal distribution.

Finally, we use a transform similar to AdaIN (Huang and Belongie 2017) that is widely applied in style transfer to produce the augmented amplitude features with rich styles as

$$\mathcal{A}_i = \gamma(\mathcal{A}_i) \frac{\mathcal{A}_i - \mu(\mathcal{A}_i)}{\sigma(\mathcal{A}_i)} + \beta(\mathcal{A}_i). \tag{7}$$

Online Update of Phase Features The phase features encode rich semantic information that is domain-invariant, and this is helpful to improve the generalization capability of the model when no semantic shift occurs. Otherwise, we need to further learn cross-domain-invariant features since the targets with different semantic categories need to be projected onto the same foreground label space in UVOS.

To learn cross-domain-invariant features, we first define the n-domain phase features as $\mathcal{P}^n = \{\mathcal{P}^1_i, \ldots, \mathcal{P}^n_i\}$, and then we learn the network $f_{\boldsymbol{\theta}}(\cdot)$ with learnable parameters $\boldsymbol{\theta}$ that projects $\mathcal{P}$ onto the same foreground label space $\mathcal{Y}$ that satisfies $f_{\boldsymbol{\theta}}(\mathcal{P}^1_i) = f_{\boldsymbol{\theta}}(\mathcal{P}^2_i) = \ldots = f_{\boldsymbol{\theta}}(\mathcal{P}^n_i) \in \mathcal{Y}$. Finally, we average the n output features to construct more robust cross-domain-invariant transformation as

$$\bar{f}_{\boldsymbol{\theta}}(\mathcal{P}^n) = \frac{\sum_{k=1}^{n} f_{\boldsymbol{\theta}}(\mathcal{P}^k_i)}{n} = (1-\rho)\bar{f}_{\boldsymbol{\theta}}(\mathcal{P}^{n-1}) + \rho f_{\boldsymbol{\theta}}(\mathcal{P}^n_i), \tag{8}$$

where the parameter $\rho = 1/n$.

The above online update process in (8) is an EMA operator (Klinker 2011) which can effectively smooth cross-domain-variant noisy information while keeping the foreground features of different semantic categories consistent (see Figure 4), and this helps to well handle the *"semantic shift"* issue in UVOS.

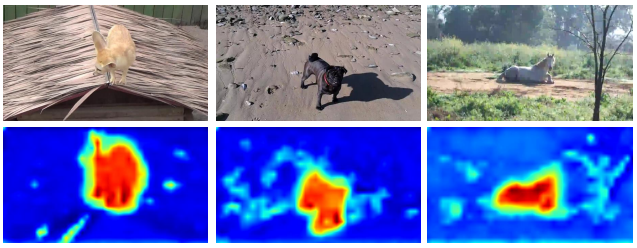

Figure 4: Visualization of the updated phase features. The online phase update strategy (8) projects *fox*, *dog*, and *horse* semantic features onto the same foreground label space, which helps to learn cross-domain-invariant features that can reduce *"semantic shift"* challenge.

Loss Function

The loss $\mathcal{L}$ is a combination of cross-entropy (CE) loss $\mathcal{L}_{CE}$ and Intersection-over-Union (IoU) loss $\mathcal{L}_{IoU}$ (Ren et al. 2021), over the predicted segmentation mask $\mathbf{P} \in [0,1]^{H \times W}$ and ground-truth mask $\mathbf{G} \in \{0,1\}^{H \times W}$. The loss $\mathcal{L}$ is defined as follows

$$\mathcal{L} = \lambda \mathcal{L}_{CE} + (1-\lambda)\mathcal{L}_{IoU}, \tag{9}$$

where $\mathcal{L}_{CE} = -\sum_{i,j} \mathbf{G}_{ij} log \mathbf{P}_{ij}$, $\mathcal{L}_{IoU} = 1 - \frac{\min(\mathbf{G}_{ij}, \mathbf{P}_{ij})}{\max(\mathbf{G}_{ij}, \mathbf{P}_{ij})}$, and λ is a trade-off coefficient.

Experiments

Implementation Details

We apply the SegFormer (Xie et al. 2021) weights pretrained on the ImageNet dataset to initialize the weights of our GFA. The training set consists of two parts: (a) all the training data in DVAIS-2016 (Perazzi et al. 2016), which contains 30 videos with about 2,000 frames. (b) a subset of 10,000 frames are selected from YouTubeVOS-2018 (Xu et al. 2018) using one frame every 10 frames sampling strategy. Our experiments follow the common practices as in (Zhang et al. 2021; Zhou et al. 2020) with training on (a) and (b) and fine-tuning on (b). All images are resized to $512 \times 512 \times 3$ pixels, and the RAFT (Teed and Deng 2020)

Methods	Aeroplane	Bird	Boat	Car	Cat	Cow	Dog	Horse	Motorbike	Train	Average
AGS	87.7	76.7	72.2	78.6	69.2	64.6	73.3	64.4	62.1	48.2	69.7
AGNN	71.1	75.9	70.7	78.1	67.9	69.7	77.4	67.3	68.3	47.8	70.8
COSNet	81.1	75.7	71.3	77.6	66.5	69.8	76.8	67.4	67.7	46.8	70.5
MATNet	72.9	77.5	66.9	79.0	73.7	67.4	75.9	63.2	62.6	51.0	69.0
GraphMem	86.1	75.7	68.6	82.4	65.9	70.5	77.1	72.2	63.8	47.8	71.4
WCS-Net	81.8	81.1	67.7	79.2	64.7	65.8	73.4	68.6	69.7	49.2	70.5
AMCNet	78.9	80.9	67.4	82.0	69.0	69.6	75.8	63.0	63.4	57.8	71.1
RTNet	84.1	80.2	70.1	79.5	71.8	70.1	71.3	65.1	64.6	53.3	71.0
HFAN	84.7	80.0	72.0	76.1	76.0	71.2	76.9	71.0	64.3	61.4	73.4
TMO	85.7	80.0	70.1	78.0	73.6	70.3	76.8	66.2	58.6	47.0	71.5
GFA (ResNet)	82.9	<u>81.2</u>	<u>73.5</u>	80.9	<u>78.9</u>	68.8	75.6	<u>71.5</u>	64.2	<u>58.1</u>	<u>73.6</u>
GFA (Segformer)	<u>87.2</u>	**85.5**	**74.7**	**82.9**	**80.4**	**72.0**	**79.6**	67.8	61.3	55.8	**74.7**

Table 1: Quantitative results for each category and overall average on Youtube-objects dataset, in which the top two performances are expressed in "bold" and "underline" fonts, respectively.

Methods	DAVIS-2016			FBMS
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$
COSNet	80.0	80.5	79.4	75.6
MATNet	81.6	82.4	80.7	76.1
DFNet	82.6	83.4	81.8	-
3DC-Seg	84.5	84.3	84.7	71.5
FSNet	83.3	83.4	83.1	-
F2Net	83.7	83.1	84.4	77.5
TransportNet	84.8	84.5	85.0	78.7
AMCNet	84.6	84.5	84.6	76.5
RTNet	85.2	85.6	84.7	-
CFANet	82.8	83.5	82.0	-
IMPNet	85.6	84.5	86.7	77.5
DBSNet	85.3	85.9	84.7	78.5
HFAN	<u>86.7</u>	<u>86.2</u>	<u>87.1</u>	76.1
TMO	86.1	85.6	86.6	79.9
GFA (ResNet)	86.3	85.9	86.7	<u>80.1</u>
GFA (Segformer)	**88.2**	**87.4**	**88.9**	**82.4**

Table 2: Quantitative results on the DAVIS-2016 and FBMS test sets, and the top three performances indicated in "bold" and "underline" fonts, respectively. For FBMS, we only report metric $\mathcal{J}$.

is adopted to estimate optical flow. The data augmentation strategies includes random rotation, random horizontal flip, random cropping and color enhancement during training. We use four NVIDIA 2080TI GPUs with 4 batch size in each GPU, and the total batch size is set to 16. During training, the model is optimized using the AdamW optimizer (Loshchilov and Hutter 2018) with a cosine decay schedule. The initial learning rate and weight decay are set to 1e-4 and 1e-4 respectively. The trade-off factor λ in the segmentation loss is set to 0.5, and directly yield the binary segmentation mask without any post-processing technique.

Datasets and Evaluation Metrics

We conduct extensive evaluations on three benchmarks including DAVIS-2016 (Perazzi et al. 2016), FBMS (Ochs, Malik, and Brox 2013), and Youtube-objects (Prest et al. 2012). We report the evaluation results in terms of region similarity $\mathcal{J}$, boundary accuracy $\mathcal{F}$ and overall average $\mathcal{J}\&\mathcal{F}$.

Comparison with the State-of-The-Arts

Table 1 lists the quantitative results of our GFA for each category and overall average against the state-of-the-art methods in terms of $\mathcal{J}$ on Youtube-objects (Prest et al. 2012). From the results, we can notice that our GFA obtains the best performance on the overall average and six different semantic categories such as bird and boat. In addition, there are many cases of out-of-distribution scenarios on Youtube-objects, where our GFA still achieves state-of-the-art performance. This adequately demonstrates that our GFA method can learn robust domain-invariant features to improve the generalization ability of the model.

Table 2 lists the comparison results of our GFA against the state-of-the-arts on DAVIS-2016 (Perazzi et al. 2016) and FBMS (Ochs, Malik, and Brox 2013). We can observe from the results that our GFA yields the best performance compared to the existing state-of-the-art UVOS methods in all evaluation metrics. Specifically, our GFA outperforms HFAN 1.4% and 6.3% with a significant margin in terms of $\mathcal{J}\&\mathcal{F}$ on DAVIS-2016 as well as in terms of $\mathcal{J}$ on FBMS, respectively. This verifies the proposed GFA can reduce the dependency of the model on the training data and enhance the generalizability.

Ablation Study

In this section, we investigate the contribution of each component by experiments. We employ standard metrics to measure performance on DAVIS-2016 and FBMS.

Impact of transformer as baseline. As shown in Section , we adopt the transformer-based model as a baseline. To prove its validity, we replace it with ResNet101 (He et al.

Module Variants			DAVIS-2016	FBMS	
Res	Segf	A-Samp	P-EMA	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$
✓				83.5	77.5
	✓			85.3	79.2
✓		✓		85.8	79.7
	✓	✓		87.5	81.0
✓		✓	✓	86.3	80.1
	✓	✓	✓	88.2	82.4

Table 3: Ablation study on DAVIS-2016 and FBMS for different variants of the proposed GFA method, in which we replace SegFormer with ResNet101 to evaluate the impact of each individual component in our GFA method. "Res", "Segf", "A-Samp" and "P-Enh" denote ResNet101, SegFormer, sampling on amplitude features and the online update of phase features, respectively.

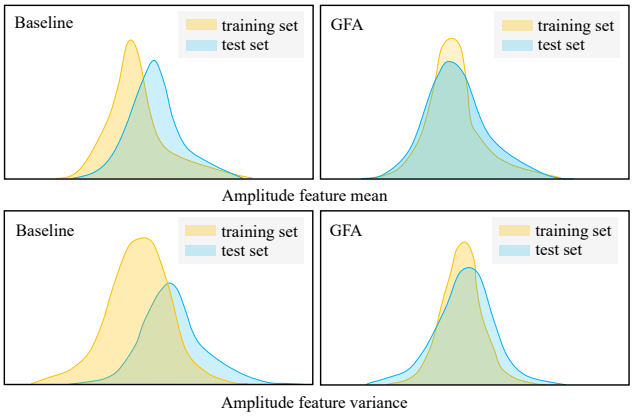

Figure 5: Visualization of the distribution of the amplitude feature statistics between the training set and testing set.

2016) and obtain a CNN-based model and the results are reported in Table 3. From the results, we have the following observations: (1) The transformer-based model outperforms the CNN-based model on both data sets. This may attribute to the fact that the transformer applies less inductive bias, leading to better results. (2) Our GFA is a plug-and-play module that is suitable to both the transformer-based and the CNN-based models and generalizes well to test videos.

Impact of amplitude feature augmentation. Now we augment amplitude features via Gaussian sampling to the transformer-based baseline, and the result is reported in the fourth row of Table 3. As seen, the proposed GFA achieves 2.2% absolute improvement in terms of $\mathcal{J}\&\mathcal{F}$ on DAVIS-2016 and 2.3% improvement in terms of $\mathcal{J}$ on FBMS. Furthermore, Figure 5 shows the amplitude feature statistics distribution between training and test sets. As seen, the distribution of our GFA method is closer than the baseline about the mean and variance of the feature statistics. These all verify that amplitude feature augmentation in GFA can produce a variety of style features to reduce the model overfitting and improve the generalizability.

Model	IIW	DAVIS16		FBMS
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$
Base w/ SegF	5.2E-06	84.8	85.7	79.2
GFA w/ SegF	2.8E-06	87.4	88.9	82.4

Table 4: Results on DAVIS16 and FBMS test sets with and without the proposed GFA method, and we also report the IIW metric to measure the generalization of the model.

Impact of the phase feature update. we perform ablation experiments to prove the significance of the online update of phase features in UVOS, and the result can be seen in Table 3. As seen, without the phase enhancement method, the performance of the model drops about 1% in terms of all evaluation metrics on both datasets (row 4 vs. row 6). This indicates that the online update process can learn cross-domain-invariant features to improve the generalization of the model, leading to better performance.

Impact of generalizable Fourier augmentation. We also add the generalizable Fourier augmentation method to the baseline and report the results in the last row of Table 3. As seen, our GFA method improves the results by 2.9% in terms of $\mathcal{J}\&\mathcal{F}$ on DAVIS-2016 and by 3.2% in terms of $\mathcal{J}$ on FBMS (row 2 vs. row 6). Our proposed method can achieve significant improvements over all evaluation metrics on both datasets. In the second column of Table 4, our GFA method achieves a lower generalization gap (lower IIW) than the transformer-based baseline. These all demonstrate that the proposed GFA could learn domain-invariant features to improve the generalization performance.

Conclusion

This paper has presented a GFA framework for UVOS, which consists of a Gaussian sampling and an online update designs, to improve the model's generalizability. Specifically, the amplitude features of image Fourier Transform are sampled from a Gaussian distribution to produce diverse scene style augmentation features, which can reduce the dependency of the model on the training set and alleviate *"sence shift"* issue. To further improve the generalization, the semantic-rich phase features are online updated via EMA by updating the phase features from the past frames, which can learn cross-domain-invariant features to reduce *"semantic shift"*. Extensive experiments on seven challenging benchmarks have verified that our GFA achieves the favorable performance against the state-of-the-art methods in terms of all evaluation metrics.

Acknowledgments

This work is supported in part by the NSFC under Grants 62276141, 61872189, U20B2065 and 62272468.

References

Brigham, E. O.; and Morrow, R. 1967. The fast Fourier transform. *IEEE spectrum*.

Chen, X.; Lan, X.; Sun, F.; and Zheng, N. 2020. A boundary based out-of-distribution classifier for generalized zero-shot learning. In *ECCV*.

Cho, S.; Lee, M.; Lee, S.; Park, C.; Kim, D.; and Lee, S. 2023. Treating Motion as Option to Reduce Motion Dependency in Unsupervised Video Object Segmentation. In *WACV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Huang, J.; Guan, D.; Xiao, A.; and Lu, S. 2021. Fsdr: Frequency space domain randomization for domain generalization. In *CVPR*.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*.

Jeon, S.; Hong, K.; Lee, P.; Lee, J.; and Byun, H. 2021. Feature stylization and domain-aware contrastive learning for domain generalization. In *MM*.

Ji, G.-P.; Fu, K.; Wu, Z.; Fan, D.-P.; Shen, J.; and Shao, L. 2021. Full-duplex strategy for video object segmentation. In *ICCV*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Klinker, F. 2011. Exponential moving average versus moving exponential average. *Mathematische Semesterberichte*.

Li, X.; Dai, Y.; Ge, Y.; Liu, J.; Shan, Y.; and DUAN, L. 2022. Uncertainty Modeling for Out-of-Distribution Generalization. In *ICLR*.

Loshchilov, I.; and Hutter, F. 2018. Fixing weight decay regularization in adam.

Ochs, P.; Malik, J.; and Brox, T. 2013. Segmentation of moving objects by long term video analysis. *TPAMI*.

Oliva, A.; and Torralba, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*.

Pei, G.; Shen, F.; Yao, Y.; Xie, G.-S.; Tang, Z.; and Tang, J. 2022. Hierarchical Feature Alignment Network for Unsupervised Video Object Segmentation. In *ECCV*.

Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Van Gool, L.; Gross, M.; and Sorkine-Hornung, A. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*.

Prest, A.; Leistner, C.; Civera, J.; Schmid, C.; and Ferrari, V. 2012. Learning object class detectors from weakly annotated video. In *CVPR*.

Ren, S.; Liu, W.; Liu, Y.; Chen, H.; Han, G.; and He, S. 2021. Reciprocal transformations for unsupervised video object segmentation. In *CVPR*.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*.

Tobin, J.; Fong, R.; Ray, A.; Schneider, J.; Zaremba, W.; and Abbeel, P. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *IROS*.

Tokmakov, P.; Alahari, K.; and Schmid, C. 2017. Learning video object segmentation with visual memory. In *ICCV*.

Wang, Z.; Huang, S.-L.; Kuruoglu, E. E.; Sun, J.; Chen, X.; and Zheng, Y. 2022. PAC-Bayes Information Bottleneck. In *ICLR*.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*.

Xu, N.; Yang, L.; Fan, Y.; Yang, J.; Yue, D.; Liang, Y.; Price, B.; Cohen, S.; and Huang, T. 2018. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV*.

Xu, Q.; Zhang, R.; Zhang, Y.; Wang, Y.; and Tian, Q. 2021. A fourier-based framework for domain generalization. In *CVPR*.

Yang, Y.; and Soatto, S. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *CVPR*.

Yue, Z.; Wang, T.; Sun, Q.; Hua, X.-S.; and Zhang, H. 2021. Counterfactual zero-shot and open-set visual recognition. In *CVPR*.

Zhang, K.; Zhao, Z.; Liu, D.; Liu, Q.; and Liu, B. 2021. Deep transport network for unsupervised video object segmentation. In *ICCV*.

Zhou, T.; Wang, S.; Zhou, Y.; Yao, Y.; Li, J.; and Shao, L. 2020. Motion-attentive transition for zero-shot video object segmentation. In *AAAI*.