Stable Autofocus with Focal Consistency Loss

Sangwon Lee* Myungsub Choi Nagyeong Lee Hyong-euk Lee
Samsung Advanced Institute of Technology (SAIT)
gobiree@gmail.com {ms00.choi, nxxbiii.lee, hyongeuk.lee}@samsung.com

Abstract

Autofocus aims to accurately position the camera lens to bring the desired region of interest into focus. Conventional works search for the sharpest frame within the lens movement. However, sharpness measure in many real-world settings is ambiguous and may cause a focus hunting problem, where the lens continuously moves back and forth to search for the accurate position. To mitigate this problem, we introduce a simple yet powerful loss function, specifically designed to produce consistent outputs in autofocus systems. The proposed Focal Consistency Loss (FCL) allows autofocus models to better learn the geometric cues relative to each initial position of the lens, significantly reducing distracting lens movement and enhancing the user experience when taking a photo. Furthermore, we improve autofocus stability by utilizing multiple consecutive frames in a practical way. Experimental results show the effectiveness of FCL in various practical scenarios, including multi-frame autofocus for both conventional and dual-pixel images.

1. Introduction

Focusing on the target subject is typically the first action for a photographer to take before capturing a photo. A good initial point would be using the autofocus (AF) feature of the camera, since modern cameras already provide sophisticated software/hardware support for rapidly moving the lens to search for the position that leads to a sharp in-focus scene. However, AF system can sometimes struggle to lock onto a subject, resulting in the lens constantly shifting in and out of focus. This problem is called *focus hunting* (also known as *lens* hunting), where the lens "hunts" for the focusing position but is unable to decide on the correct point. Focus hunting tends to occur when the scene lacks sufficient contrast or when the lighting conditions are poor. This issue can greatly impact the user experience, as it often results in a completely blurry, out-of-focus image that fails to capture the scene effectively.

*Work done while at SAIT.

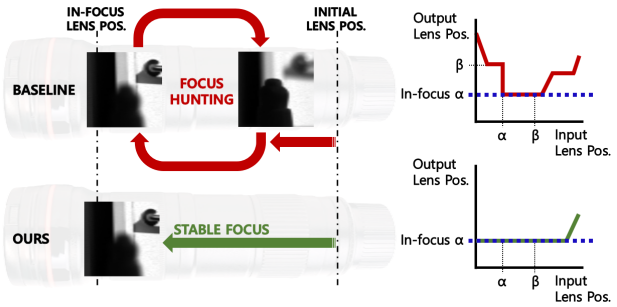

Figure 1. Motivation of our work: Autofocus models can struggle to find the in-focus lens position when the scene is complex, such as when there are multiple objects at varying depths. This may result in AF systems not converging and encounter distracting lens movements ($\alpha \leftrightarrow \beta$) called *focus hunting*. In this work, we propose a focal consistency loss with multi-frame approach and achieve significantly better stability, even for difficult cases that the state-of-the-art AF model (BASELINE) fails.

Most of the existing solutions for focus hunting resort to heuristic guidelines for photographers, such as adjusting the camera settings for better shooting condition or simply switching to manual focus [12]. The hunting problem still exists even for the latest learning-based AF approaches [4, 13], which demonstrated substantial improvements in AF performance compared with conventional AF algorithms. We show a motivating example in Fig. 1, where the state-of-the-art AF method [4] encounters focus hunting for a challenging scene, while our proposed approach successfully mitigates the cyclic lens wobbling.

In this paper, we present two methodologies to make the AF models more stable. First, we introduce a novel loss function, Focal Consistency Loss (FCL), to enforce consistent predictions for an AF model. Specifically, we reorganize the training samples so that multiple patches from the same scene but with different initial lens positions belong to a single mini-batch. For each scene, note that the ground truth focus point should be consistent regardless of the initial lens position. Therefore, our proposed FCL is implemented to enforce the intra-scene (same scene, different initial lens positions) predictions to be convergent, which can improve the robustness of the model without any addi-

tional computational burden at inference time.

Second, we explore a practical multi-frame AF problem setting to reliably infer the focus point. Notably, we use frames from consecutive lens positions, which could be easily (and quickly) obtained when the lens is moving towards its estimated focusing position. While providing a further boost in AF accuracy and stability, our multi-frame model exhibits negligible extra computational complexity, because we only modify the first convolutional layer of our AF network to receive additional input channels. Training the multi-frame model together with FCL enables our AF model to observe diverse intra-scene examples with the same ground truth label (focusing lens position). Consequently, our AF model could better learn the meaningful geometric cues and patterns, which leads to a more accurate and stable autofocus. In experiments, we show that the proposed FCL and the multi-frame scheme could stabilize various existing AF models and greatly reduce the focus hunting issue in practice.

Our contributions can be summarized as follows:

- We propose Focal Consistency Loss, a novel loss function that minimizes the deviations of intra-scene predictions and reduce focus hunting.
- We introduce a practical AF problem setting that use multiple frames and achieve state-of-the-art results.

2. Related Work

2.1. Autofocus

Existing methods for camera autofocus can be categorized into two main streams: contrast-detection autofocus (CDAF) and phase-detection autofocus (PDAF). CDAF aims to find the position of the lens which gives the sharpest image using a pre-defined contrast metric. Searching for the peak position is typically done by hill-climbing, and the contrast metrics are computed by image statistics [9, 18, 37] or frequency-domain representations [17, 20, 22, 32, 36]. However, the target contrast metric can be noisy and may consist of multiple local maxima within the range of lens movement. In such cases, the estimated focusing points can be different depending on the initial lens position, since the CDAF algorithm would converge to the nearest local maximum on the sharpness curve instead of giving consistent results across the full range. To improve the stability of CDAF algorithms, numerous efforts have been proposed, including more a robust contrast metric [11, 23, 34] or better contrast measurements over multiple frames [10, 35].

On the other hand, PDAF algorithms calculate the disparity between the left/right dual-pixel data and converts the disparity into a focus distance using a pre-computed calibration map. While able to predict the lens position in one shot, PDAF algorithms are susceptible to errors from geometric distortions or low-light noise [4, 19, 29], which may lead to a focus hunting problem.

Recently, deep learning-based approaches have been proposed and greatly improved the overall AF performance for both CDAF and PDAF [3, 4, 13, 15, 31]. In particular, Hermann *et al*. [13] first introduced a large-scale dataset for AF and outperformed all existing AF approaches, and Choi *et al*. [4] further improved the dual-pixel AF performance. However, both methods [4, 13] did not consider the stability of AF models. We assert that learning-based AF models with the state-of-the-art performance are still prone to focus hunting problems (as we illustrate in Fig. 1) and introduce a novel focal consistency loss to mitigate the issue.

In addition to the image AF, efforts dedicated to improving the smoothness of *video* AF are also related to our work. Tsai and Chen [30] tried to mitigate the *bouncing* lens movements by using a Kalman filter to estimate more accurate lens positions, but they show limited lab-scale experimental results. Abuolaim *et al*. [1] performed a more thorough user study to demonstrate the effectiveness of their bidirectional LSTM module with a weighted moving average for smooth lens movements. Our work is different from Abuolaim *et al*. [1] in that we can enforce prediction consistency within a single scene at a single time step. Also, our method is model-agnostic and can be applied on top of many different AF models including [1].

In this work, we use [4] and [13] as our baseline models, extend them to utilize multiple consecutive input frames, and improve their stability while preserving accuracy by training with the proposed focal consistency loss.

2.2. Consistency Regularization Loss

Consistency regularization technique is widely used in semi-supervised learning to make the model more robust to semantic-preserving perturbations while leveraging unlabeled data [2, 16, 21, 27, 33, 38]. This can help the model to learn robust and discriminative features as well as to reduce overfitting to the limited labeled data. Existing works on consistency regularization are implemented by matching multiple output values obtained from input data augmentation [5–7, 39], dropout [26], or random max-pooling layers [21, 27]. The regularization is performed by minimizing the mean squared error between the output predictions or Kullback-Leibler (KL) divergence between the output probability distributions. While our work follows the main philosophy of consistency regularization, we do not rely on random perturbations like data augmentation but use the intrinsic data characteristic of the AF problem. Also, our method is fully supervised, and our proposed consistency loss is motivated to oversee a physically meaningful feature of defocus blur. There exists a similar recent approach in natural language processing domain that introduces a consistency loss based on Wasserstein distance between simi-

lar documents for offensive text detection [24]. However, to the best of our knowledge, our work is the first to explore and adapt the consistency regularizing scheme to the new domain of low-level computer vision problems.

3. Methods

To formulate the autofocus problem, we follow the previous learning-based works [4, 13]. First, we quantize the continuous lens positions into n discrete focus distances and denote the index as $f \in \{1, \dots, n\}$. Let $\{I_f\}$ represent the set of input patches. We term the individual patch I_f as a *focal slice*, the full collection of patches obtained at different focus distances $\{I_f \mid f \in 1, \dots, n\}$ as a *focal stack*, and f as a *focal index*. We assume that the region-of-interest (RoI) to focus is already selected by the user, so that I_f indicates the image *patch* cropped from the field-of-view.

We model AF as a classification problem, where the goal is to accurately predict the ground truth (GT) focal index f_*. However, unlike a typical classification setting where each class is independent, the GT focal index for AF is closely related to the neighboring focal indices, which is why existing works [4, 13] use an ordinal regression loss [8] for training. This encourages the model to give lower loss for the nearby focal indices compared with a distant index.

In the current problem formulation, the AF model is trained with a *single* GT focal index supervision of f_*. This is because the lens position should be fixed for a certain time step, and we cannot have multiple focus distance at the same time. However, there exists confusing cases with mixed depths, and the true focus depends on the user's intention. For instance, focusing on the foreground object may have the same sharpness metric as focusing on the background, where some users may want to focus on the foreground object but other users think of it as occlusion and want to focus on the background. Even in such cases with multi-modal distribution of sharpness metrics, we assume in this work that there is only a single (more dominant) mode, which is determined by the GT labels in the training dataset. This setting is in line with prior research, allowing for fair experimental comparisons.

In this work, we introduce two new modifications to improve the AF stability: Focal Consistency Loss (FCL) and consecutive multi-frame inputs. FCL is included during training as an additional loss function that stimulates intra-scene convergent behavior, and multi-frame inputs allow for more stable predictions. Note that our method is model-agnostic and can be easily applied to the other AF baselines. We describe the details in the following subsections.

3.1. Focal Consistency Loss

Given a specific ground truth focal index f_*, the target distribution of soft labels y for ordinal regression [8] can be computed by

$$y_i = \frac{\exp(-|f_i - f_*|^2)}{\sum_{j=1}^{n} \exp(-|f_j - f_*|^2)}, \quad \forall i \in \{1, \dots, n\}, \quad (1)$$

and the ordinal regression loss is calculated as the cross entropy between the soft label distribution y and the softmax output probability p of our AF model:

$$\mathcal{L}_{\text{ORD}} = -\sum_{i=1}^{n} y_i \log(p_i), \quad p_i = \frac{\exp(x_i)}{\sum_{j=1}^{n} \exp(x_j)}, \quad (2)$$

where x_i is the output logit for the i-th class before the softmax function.

Ideally, our AF model should be able to predict consistent focal index regardless of the initial focus distance, *i.e.* predictions for input patches I_0, I_1, ..., and I_n should all be the same with the GT focal index f_*. However, this is usually not the case in practice; while ordinal regression loss is useful in leveraging the ordinal characteristic of the focal indices, it cannot enforce multiple predictions from different initial focal indices to be consistent.

To explicitly supervise our AF model to give homogeneous outputs, we propose Focal Consistency Loss (FCL) and suggest two different ways for realization: symmetric Kullback-Leibler divergence (KLD) and mean standard deviation (MSD). Both implementations drive the intra-scene predictions to be more consistent, as illustrated in Fig. 2.

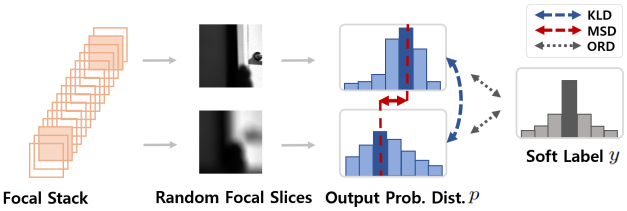

Figure 2. Overview of the proposed Focal Consistency Loss (FCL). For different focal slices from the same scene, the output probability distribution may vary. Our FCL ensures consistent predictions by minimizing the KL-Divergence (KLD) and mean standard deviation (MSD) of the intra-scene outputs.

Symmetric KL-Divergence (KLD). Given a single scene, consider two distinct input focal indices, f_i and f_j (where $i \neq j$), and the corresponding output probabilities for each focal index as $p(f_i)$ and $p(f_j)$. Our goal is to make the two distributions $p(f_i)$ and $p(f_j)$ to be similar, to enforce consistent intra-scene predictions; if all distributions $p(f_i), \forall i \in \{1, \dots, n\}$ are the same, then our AF model predictions are perfectly consistent for the given scene, and there would be no focus hunting or other distracting lens movements. To achieve this goal, we use KL-Divergence, which is widely adopted for measuring the difference between two probability distributions. However, since KL-Divergence is an asymmetric metric, we use its symmetric

form to implement the focal consistency loss:

$$\mathcal{L}_{\text{KLD}} = \frac{1}{2}\left(D_{\text{KL}}\left(p(f_i) \parallel p(f_j)\right) + D_{\text{KL}}\left(p(f_j) \parallel p(f_i)\right)\right), \tag{3}$$

where we denoted the KLD-type focal consistency loss as $\mathcal{L}_{\text{KLD}}$, and D_{KL} stands for the standard KL-Divergence, which is computed as:

$$D_{\text{KL}}\left(p(f_i) \parallel p(f_j)\right) = \sum_{k=1}^{n} p_k(f_i) \log \frac{p_k(f_i)}{p_k(f_j)}. \tag{4}$$

Note that $\mathcal{L}_{\text{KLD}}$ does not depend on the ground truth focal index f_* and only considers the consistency between our model outputs within each scene. For each training iteration, the final loss is calculated by averaging the KLD values for all scenes in the mini-batch. When there are more than two focal indices from the same scene, we compute the KLD for all pairs; in practice, we randomly choose 4 input focal indices for each scene, resulting in $\binom{4}{2} = 6$ pairs.

Mean Standard Deviation (MSD). An alternative approach of measuring the focal consistency is to directly compute the deviations for each intra-scene prediction of $p(f_i)$. We can compute the standard deviation of all predictions for a given scene as:

$$\mathcal{L} = \text{STD}\left(\{\text{argmax}_j\, p_j(f_i) | \; i = 1, \ldots, F\}\right), \tag{5}$$

where STD is the standard deviation operator and F is the number of input focal indices from the same scene. Note that STD is a set metric (unlike KLD, which is a pairwise metric) and can be easily computed for input focal indices $F \geq 2$. This allows us to utilize Eq. (5) as an *evaluation metric that can measure the focal consistency* by computing it across the full focal stack ($F = n$). However, since argmax is not differentiable, it cannot be used as a loss function. To address this limitation, we propose to use a Soft-argmax function, which is a smooth and differentiable approximation of the one-hot argmax:

$$\text{soft argmax}(x) = \sum_{i=1}^{n} \frac{i}{n} \frac{\exp(\beta x_i)}{\sum_{j=1}^{n} \exp(\beta x_j)}, \tag{6}$$

where x_i is the output logit for the i-th class, and β is a temperature parameter [14] that controls the sharpness of the softmax distribution (we use $\beta = 2.0$ in practice). Our final loss, named MSD as we take the mean value of the standard deviations for all scenes, is then calculated by the following equation:

$$\mathcal{L}_{\text{MSD}} = \text{STD}(\{\text{soft argmax}(x_i) \mid i = 1, \ldots, F\}). \tag{7}$$

In practice, we take the weighted sum of the KLD and MSD losses to compute our final FCL:

$$\mathcal{L}_{\text{FCL}} = \lambda_{\text{KLD}}\mathcal{L}_{\text{KLD}} + \lambda_{\text{MSD}}\mathcal{L}_{\text{MSD}}, \tag{8}$$

where λ_{MSD} and λ_{KLD} are hyperparameters that control the trade-offs between the accuracy ($\mathcal{L}_{\text{ORD}}$) and the consistency ($\mathcal{L}_{\text{FCL}}$) metrics. The final training loss function of our AF model is the sum of the original ordinal loss and the proposed focal consistency loss, which is computed as:

$$\mathcal{L} = \mathcal{L}_{\text{ORD}} + \mathcal{L}_{\text{FCL}}. \tag{9}$$

3.2. Multi-frame Input

In addition to FCL, using multiple input frames can also help stabilize the AF model performance. Previously, Herrmann *et al.* [13] proposed the multi-step problem setting for AF, where the lens moves to the first-step predicted position and then uses the two observed frames (initial input position and the first-step prediction) as the second-step AF model input. Formally, let us denote the initial lens position as f_i and the first-step model as $\mathcal{M}^{(1)}$, so that the first-step predicted lens position is $f_j = \mathcal{M}^{(1)}(f_i)$. Letting $\mathcal{M}^{(2)}$ be the second-step model, the second-step output f_k can be calculated as:

$$f_k = \mathcal{M}^{(2)}(f_i, f_j) = \mathcal{M}^{(2)}\left(f_i, \mathcal{M}^{(1)}(f_i)\right). \tag{10}$$

In this scenario, running the second-step model $\mathcal{M}^{(2)}$ requires the camera lens to move all the way to the first-step predicted position f_j. The total amount of lens movement should then be $|f_i - f_j| + |f_j - f_k|$.

On the other hand, we propose to use *consecutive* lens positions as the multi-frame inputs. For instance, computation for a two-frame setting would be $f_j = \mathcal{M}^{(1)}(f_i, f_{i+1})$, and prediction for a three-frame setting can be computed as:

$$f_k = \mathcal{M}^{(1)}(f_i, f_{i+1}, f_{i+2}). \tag{11}$$

Note that we only use the first-step model $\mathcal{M}^{(1)}$ in our multi-frame setting. We claim that this is a much more practical scenario compared with the multi-*step* setting of Herrmann *et al.* [13], since we can estimate the accurate position in one-shot, and we also do not need to keep the additional parameters for $\mathcal{M}^{(2)}$. In addition, the total amount of lens movement would be $|f_i - f_k|$, which is almost always less than or equal to the multi-*step* lens movements due to triangle inequality. In our implementation, we always obtained the consecutive input frames in increasing order: *e.g.* f_i, f_{i+1} for D2, if we start at f_i. Thus, there might be some cases where our multi-frame model need to move the lens more than the multi-step settings if $f_k \leq f_i$, but such slight lens movements are often negligible in terms of runtime, compared to AF processing. For additional training/implementation details, please refer to our supplementary document.

4. Experiments

We perform extensive experiments for various different problem settings. Our notation for each setting follows Her-

rmann *et al*. [13] and summarized as follows:

- D1 $\sim$ D5 represent using $1 \sim 5$ input frames, where each frame is a 2-channel image consisting of the left and right dual-pixel RAW images.
- I1 $\sim$ I5 represent using $1 \sim 5$ input frames, where each frame is a conventional single-channel RAW image.
- D* and I* represent the full-stack performance for dual-pixel and conventional input images, respectively. This setting is shown only for comparison purposes as our multi-frame performance upper bounds.

Dataset. Following the previous works [4, 13], we use the large-scale AF dataset proposed by Herrmann *et al*. [13]. The dataset includes 510 scenes and 49 focal depths. We use the input patches of 128×128 resolution, resulting in 387,000 patches for training and 56,800 for testing. We report the performance on the test patches. For additional details on the dataset, we refer the readers to [13].

Baselines. We use two baseline models from recent works [4, 13]. For Choi *et al*. [4], we use the MobileNet-v2 [28] based model and denote it as *AFPE*. For Herrmann *et al*. [13], the network architecture is also MobileNet-v2, but the channel width for each layer is multiplied by 4, as stated in [13]. We denote this model as *L2A*. Note that both baselines are 49-class classification models.

To reduce notation clutter, we denote the model trained with $\mathcal{L}_{\text{FCL}}$ of Eq. (9) as $\{\texttt{BASELINE}\}$+$\{\texttt{LOSS}\}$. For instance, an L2A model trained with $\mathcal{L}_{\text{FCL}}$ is called as L2A+FCL ($\lambda_{\text{KLD}} \neq 0, \lambda_{\text{MSD}} \neq 0$ in Eq. (8)). An AFPE model trained with $\lambda_{\text{MSD}} = 0$ is denoted AFPE+KLD, and AFPE+MSD for $\lambda_{\text{KLD}} = 0$, likewise.

Evaluation Metrics. We follow the same evaluation protocol as the baseline methods [4, 13], while introducing two new focal consistency metrics: *MSD** and *Total Variation (TV)* [25]. MSD* measures the variability of predictions for each scene by calculating the mean standard deviation across all focal indices, which is computed by Eq. (5) with $F = n$. We denote the metric as MSD* to distinguish the notation with MSD as a loss function computed by Eq. (7).

Total Variation, on the other hand, measures the consistency of predicted focal indices across adjacent frames by calculating the sum of absolute differences between the predictions of consecutive frames:

$$TV = \sum_i \left| \text{argmax}_j \, p_j(f_{i+1}) - \text{argmax}_j \, p_j(f_i) \right|. \quad (12)$$

TV represents the amount of variation or *jumps* in the sequence of predictions. Since MSD* is computed in a similar way as the MSD loss that we use for training, one might suspect that a better MSD* with FCL is a result of overfitting to the metric using a similar loss function. We claim that this is *not* such as case and prove that MSD* is a valid metric that is not hacked by providing TV, which has a completely different functional form as the MSD or KLD loss.

We also report the existing accuracy measures: mean absolute error (MAE) and root-mean-square error (RMSE). We average the performance of all 49 starting positions of the lens to account for the accuracy variations with respect to the initial position. All accuracy metrics (MAE, RMSE) and consistency metrics (MSD*, TV) indicate better performance with lower values. By incorporating the focal consistency metrics into evaluation, we can effectively assess our model's performance and the prediction consistency across frames for each scene, allowing for gaining new insights on the stability of an AF model.

Alg.	Type	MAE	RMSE	MSD*	TV	Type	MAE	RMSE	MSD*	TV
AFPE	D1	1.760	2.855	1.338	0.895	I1	3.629	6.083	3.947	1.534
+FCL		1.735	2.744	1.070	0.691		3.506	5.902	3.356	1.246
AFPE	D2	1.656	2.593	1.161	0.606	I2	2.547	4.358	2.647	1.401
+FCL		1.577	2.466	0.968	0.488		2.491	4.210	2.204	1.188
AFPE	D3	1.542	2.421	1.073	0.496	I3	2.222	3.705	2.097	1.151
+FCL		1.522	2.350	0.884	0.376		2.110	3.519	1.779	0.928
AFPE	D4	1.516	2.351	1.003	0.420	I4	1.990	3.299	1.814	0.935
+FCL		1.434	2.279	0.868	0.319		1.958	3.202	1.557	0.790
AFPE	D5	1.456	2.236	0.954	0.368	I5	1.883	3.081	1.681	0.825
+FCL		1.431	2.226	0.833	0.290		1.801	2.959	1.484	0.703
AFPE†	D*	1.356	2.128	-	-	I*	1.550	2.399	-	-

Table 1. Quantitative results for multi-frame settings with dual-pixel (D1 $\sim$ D5) and conventional-image (I1 $\sim$ I5) using AFPE [4] baseline and the proposed FCL. The top two methods for each metric are highlighted in red and orange. We can observe that both KLD and MSD notably improves the consistency metric MSD* for all settings while preserving the accuracy. A † indicates that values are from the reference article.

Alg.	Type	MAE	RMSE	MSD*	TV	Type	MAE	RMSE	MSD*	TV
L2A	D1	2.198	3.283	1.564	1.044	I1	3.670	6.155	4.039	1.618
+FCL		2.057	3.070	1.224	0.730		3.672	6.101	3.650	1.415
L2A	D2	2.058	3.042	1.344	0.689	I2	3.613	5.892	3.561	1.507
+FCL		1.882	2.778	1.213	0.573		3.388	5.491	2.875	1.223
L2A	D3	1.913	2.841	1.292	0.577	I3	2.451	4.070	2.360	1.089
+FCL		1.798	2.716	1.081	0.429		2.403	3.885	1.984	0.930
L2A	D4	1.811	2.719	1.231	0.484	I4	2.207	3.734	2.075	0.953
+FCL		1.727	2.636	1.051	0.397		2.149	3.452	1.649	0.758
L2A	D5	1.785	2.687	1.217	0.455	I5	2.089	3.490	1.864	0.851
+FCL		1.704	2.534	0.967	0.326		1.999	3.281	1.605	0.707
L2A†	D*	1.611	2.674	-	-	I*	1.600	2.446	-	-

Table 2. Quantitative results for multi-frame settings with dual-pixel (D1 $\sim$ D5) and conventional-image (I1 $\sim$ I5) using L2A [13] baseline and the proposed FCL.

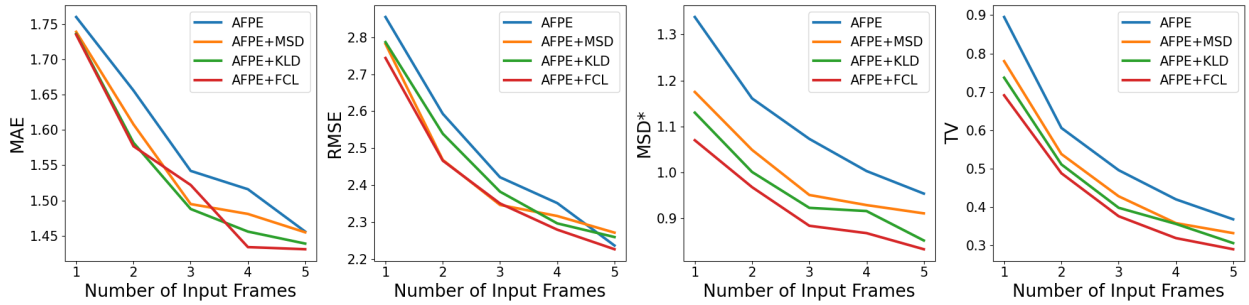

Figure 3. AF accuracy and consistency comparisons w.r.t. the number of input frames for the dual-pixel AFPE baseline for different types of FCL. The results prove the effectiveness of all types of FCL (MSD, KLD, and the combined FCL).

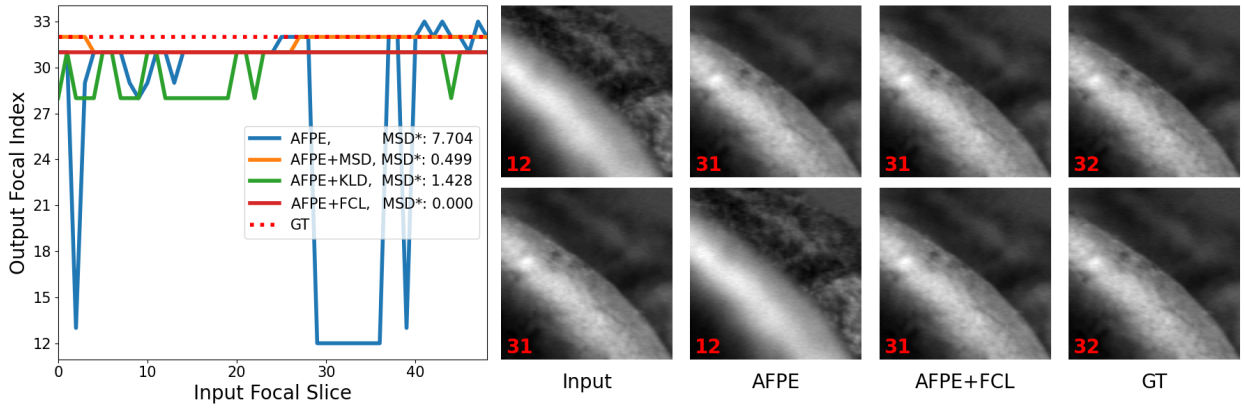

Figure 4. Qualitative comparison our proposed FCL methods with the AFPE [4] baseline for the D1 (single-slice, dual-pixel) setting. The leftmost graph illustrates the output focal index predictions for each input focal slice. The original AFPE is confused whether to focus on the foreground (index: 32) or the background (index: 12), whereas AFPE+FCL methods (including AFPE+MSD and AFPE+KLD) show more consistent predictions near the GT. Note that AFPE in this case encounters focus hunting, and the lens will oscillate between the indices 12 and 31. The red numbers in each patch indicate the focal index, with 48 being the nearest and 0 being the furthest.

4.1. Quantitative Results

In Tabs. 1 and 2, we demonstrate the quantitative results of the proposed focal consistency loss and the multi-frame schemes on AFPE and L2A baseline models, respectively, using dual-pixel inputs (D1 $\sim$ D5) and conventional input images (I1 $\sim$ I5). Since the pretrained models are not available, we reproduced and extended the model to evaluate all metrics for multi-frame settings.

For all settings across different metrics in Tab. 1, models trained with the proposed FCL outperform the corresponding baselines. In particular, FCL improves the focal consistency by 20% on the MSD* metric and 23% on the TV metric for the D1 setting, and by 15% on the MSD* metric and 19% on the TV metric for the I1 setting. Note that this achievement is done without introducing any additional computational complexity or sacrificing any AF accuracy at test time. Using multiple consecutive frames is also shown to be effective, as the focal consistency and accuracy are steadily improved as we increase the number of input frames. When it comes to the D5 setting, we argue that the AF accuracy nearly matches the full-stack performance with 0.833 MSD* and 0.290 TV, and we could observe almost no focus hunting issue at this scale except for some extreme corner cases. This presents a much more practical setting compared with the full-stack model, since capturing only 5 consecutive frames instead of the full 49 frames is significantly more efficient.

For Tab. 2, we could draw identical conclusion as for Tab. 1, which means that our FCL is model-agnostic and can be generalized across different baselines.

In Fig. 3, we visualize the accuracy (MAE and RMSE) and the focal consistency (MSD* and TV) improvements w.r.t. the number of input frames. As expected, all metrics are monotonically improved as we use more frames. We could observe similar patterns regardless of the type of our consistency loss, but the weighted combination of both MSD and KLD (AFPE+FCL) demonstrated the best performance. For more detailed quantitative results, please refer to our supplementary document.

4.2. Qualitative results

In Fig. 4, we show visual examples of how our proposed FCL can actually reduce the focus hunting issue. Specifically, for the target scene, we first draw a graph that marks

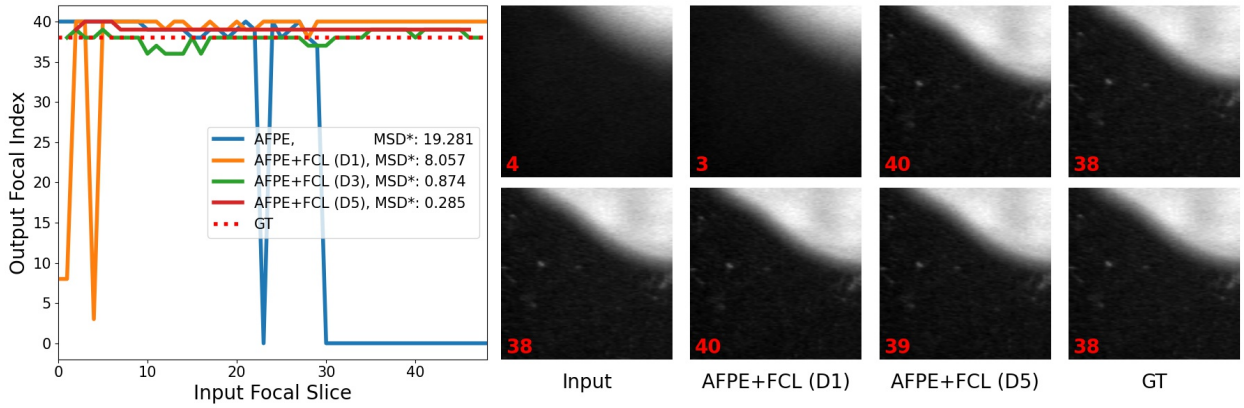

Figure 5. Qualitative comparison of results from AFPE+FCL with different multi-frame settings: D1, D3, and D5. The left graph shows that, while our FCL enhances prediction consistency, using only a single slice input (D1) may sometimes produce false answers for challenging scenes. Our multi-frame models (D3, D5) can alleviate this issue and stably converge near the GT, as illustrated in the right.

all output focal index prediction for each input focal slice, so that we can observe how the AF predictions fluctuate during the full-range lens movement. The results demonstrate that models trained with any type of FCL substantially reduce the prediction variance. In particular, if the initial lens position is at focal index 12 in Fig. 4, the original AFPE will suffer from focus hunting, while AFPE+{MSD/KLD/FCL} are all able to give more stable predictions. Note that FCL is only included during the training stage, which means the computational complexity at the inference stage is the same for all models. Thus, we can argue that the stability of any learning-based AF models can be effectively enhanced at almost no cost, by simply adding FCL to the loss function.

In Fig. 5, we show an example of the stability improvement using our multi-frame approaches. Although the D1 setting of AFPE+FCL is considerably more consistent compared with the baseline AFPE, it may still fail for difficult corner cases *e.g.* regions with little texture. However, using multiple consecutive frames can successfully reduce the prediction fluctuation with negligible additional computational complexity (see Sec. 4.3). For more qualitative results, please refer to our supplementary document (Sec. G), where we provide more diverse visual comparisons.

4.3. Analysis

Intuitively, the proposed FCL can be understood through the lens of the standard bias-variance tradeoff in machine learning. Here, *bias* represents the error between the target lens position and the model prediction, measured by MAE, while *variance* represents prediction consistency, measured by MSD*. The tradeoff between bias and variance is managed by the FCL weights λ_{KLD} and λ_{MSD}. In Fig. 6, we show an example case of how FCL affects the distribution of the model prediction to mitigate focus hunting. While this single case does not analytically guarantee the consistency, we quantitatively and qualitatively demonstrate that

Alg.	Type	MAE	RMSE	MSD*	TV	Type	MAE	RMSE	MSD*	TV
AFPE	D1	1.760	2.855	1.338	0.895	I1	3.629	6.083	3.947	1.534
+MSD		1.739	2.782	1.175	0.780		3.570	6.019	3.631	1.407
+KLD		1.736	2.787	1.130	0.737		3.533	5.898	3.471	1.285
+FCL		1.735	2.744	1.070	0.691		3.506	5.902	3.356	1.246
AFPE	D5	1.456	2.236	0.954	0.368	I5	1.883	3.081	1.681	0.825
+MSD		1.455	2.271	0.911	0.332		1.878	3.032	1.538	0.723
+KLD		1.439	2.259	0.852	0.306		1.805	3.001	1.499	0.705
+FCL		1.431	2.226	0.833	0.290		1.801	2.959	1.484	0.703

Table 3. Ablation study for with dual-pixel (D1, D5) and conventional-image (I1, I5) using AFPE [4] baseline and the MSD, KLD, FCL. FCL demonstrates the best performance in terms of both accuracy and consistency metrics.

AFPE+FCL achieves the best results *on average*, using parameters optimized to the best of our efforts. Below, we show further analysis on the empirical effects of each module and hyperparameter choice of our proposed method.

Loss Ablation. Table 3 shows the effect of the type of FCL on AF accuracy and consistency. On average, using KLD as the consistency loss performs slightly better than using MSD, and the combined FCL performs the best. This result resolves the potential problem of overfitting to MSD* by training with the MSD loss, because they have the same formulation, only different batch sizes. However, since AFPE+KLD achieves better MSD* than AFPE+MSD, we can claim that MSD* can work as a good consistency metric that is not prone to overfitting.

Multi-frame vs Multi-step. Table 4 presents a comparative analysis of the multi-frame and multi-step formulations. While D2 and 2-step settings use the same number of frames, 2-step uses twice the parameters and inference time, while D2 shows comparable MAE with negligible additional complexity. The effects of FCL is orthogonal to these settings; training with FCL consistently shows improved MSD*, and D2 is beneficial for both 1- and 2-steps.

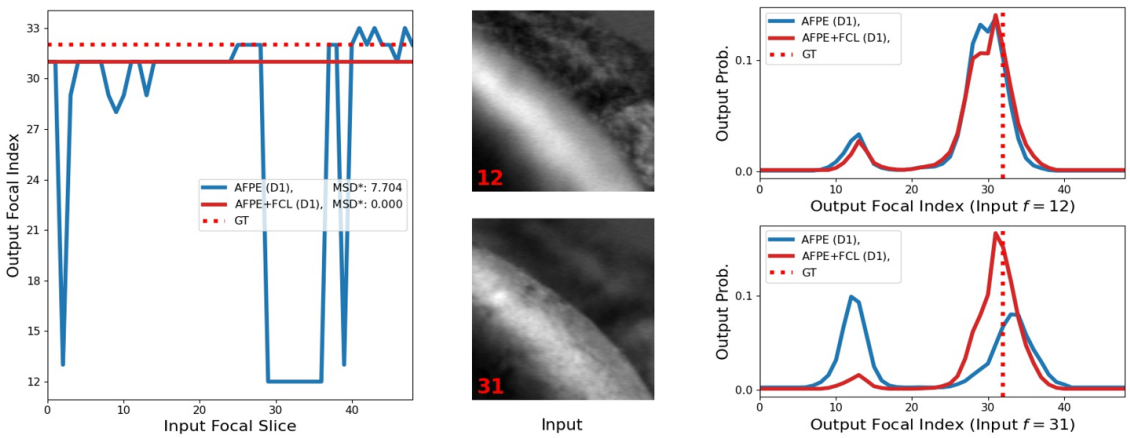

Figure 6. Example analysis on the changes of the model prediction and the output probability distribution, with respect to the different input focal slices. We show that AFPE+FCL can successfully alleviate the focus hunting issue of the original AFPE [4].

		1-step					2-step				
Alg.	Type	MAE	RMSE	MSD*	TV	Time(ms)	MAE	RMSE	MSD*	TV	Time(ms)
AFPE	D1	1.760	2.855	1.338	0.895	3.63	1.608	2.566	0.600	0.458	7.25
+FCL		1.735	2.744	1.070	0.691		1.607	2.541	0.432	0.305	
AFPE	D2	1.656	2.593	1.161	0.606	3.67	1.535	2.389	0.482	0.314	7.34
+FCL		1.577	2.466	0.968	0.488		1.491	2.330	0.356	0.210	
L2A	D1	2.198	3.283	1.564	1.044	16.41	2.100	3.120	0.794	0.622	32.81
+FCL		2.057	3.070	1.224	0.730		1.973	2.915	0.497	0.332	
L2A	D2	2.058	3.042	1.344	0.689	16.41	1.978	2.865	0.593	0.386	32.81
+FCL		1.882	2.778	1.213	0.573		1.831	2.665	0.454	0.252	

Table 4. Quantitative comparison for multi-frame (D2) and multi-step (D1, 2-step) settings using AFPE [4] and L2A [13] baselines. The D2 setting performs comparably to the 2-step setting in terms of accuracy, maintains an inference time close to D1, while the 2-step setting requires double the inference time. Runtime is measured on the Samsung Galaxy A54 device.

Calculation time. In Tab. 4, we also report the on-device inference time for each setting. In practice, AF algorithms must run locally on-device in real-time. While the current millisecond-level runtime may seem fast, there still exists multiple risk factors in real world that can lower the speed, for instance: focusing region-of-interest could be larger than 128x128, there can be other processes that simultaneously use the on-device NPU resources (*i.e.* object/face detection in the Camera App), high NPU temperature or lower clocks, *etc*. Given that modern smartphones typically support 4K-60fps video capture, the computational resources available for an AF model is extremely tight, often leaving no room for 2-step inference. To this end, the proposed multi-frame setting is practically much more beneficial in that we show negligible runtime increase as we use more number of frames, while the existing multi-step approaches require linearly-increasing time complexity.

Effects on FCL Weight. Figure 7 illustrates the results of the weight parameters w.r.t. MAE and MSD*. We found that for certain range of λ_{MSD} and λ_{KLD}, MAE remains flat or even decreases; this implies that using FCL with a properly chosen weight enables the model to improve performance by leveraging the intra-scene semantic information. However, too large values of λ_{MSD} or λ_{KLD} significantly increase the MAE and give poor AF accuracy. On the other hand, MSD* monotonically decreases *w.r.t.* increasing λ_{MSD} or λ_{KLD}, and we select the optimal value that shows the best trade-off between AF stability and accuracy. Please refer to our supplementary document (Sec. D) for the full quantitative results.

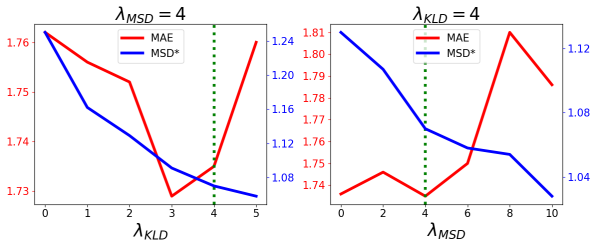

Figure 7. Effects of FCL weights λ_{KLD} and λ_{MSD} on the AF accuracy and consistency. We show the results using the AFPE baseline (D1) for each hyperparameter while fixing the other. Our selected values with optimal trade-offs are marked in dotted green.

5. Conclusion

In this work, we proposed a novel loss function, Focal Consistency Loss (FCL), designed to improve the stability of autofocus in practical real-world scenarios. FCL enabled the AF models to better understand the intra-scene geometric information, thereby enhancing both the accuracy and the prediction consistency to reduce focus hunting. In addition, we introduced and explored a practical multi-frame autofocus setting. Experimental results demonstrated the effectiveness of our proposed novelties in handling challenging practical cases. For future work, we plan to exploit the similarities between the consecutive focal slices for improved stability and reduced focus hunting.

References

[1] Abdullah Abuolaim and Michael Brown. Online lens motion smoothing for video autofocus. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020. 2

[2] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *NeurIPS*, 2014. 2

[3] Chin-Cheng Chan and Homer H. Chen. Autofocus by deep reinforcement learning. *Electronic Imaging*, 2019. 2

[4] Myungsub Choi, Hana Lee, and Hyong-euk Lee. Exploring positional characteristics of dual-pixel data for camera autofocus. In *ICCV*, 2023. 1, 2, 3, 5, 6, 7, 8

[5] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *CVPR*, 2019. 2

[6] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPRW*, 2020. 2

[7] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2

[8] Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *CVPR*, 2019. 3

[9] Judith Dijk, Michael van Ginkel, Rutger J van Asselt, Lucas J van Vliet, and Piet W Verbeek. A new sharpness measure based on gaussian lines and edges. In *10th International Conference on Computer Analysis of Images and Patterns (CAIP)*, 2003. 2

[10] Jan-Mark Geusebroek, Frans Cornelissen, Arnold WM Smeulders, and Hugo Geerts. Robust autofocusing in microscopy. *Cytometry: The Journal of the International Society for Analytical Cytology*, 2000. 2

[11] Liqiang Guo and Lian Liu. A perceptual-based robust measure of image focus. *IEEE Signal Processing Letters*, 2022. 2

[12] Damon Guy. https://www.photokonnexion.com/hunting-auto-focus-definition/. *Photokonnexion*. 1

[13] Charles Herrmann, Richard Strong Bowen, Neal Wadhwa, Rahul Garg, Qiurui He, Jonathan T. Barron, and Ramin Zabih. Learning to autofocus. In *CVPR*, 2020. 1, 2, 3, 4, 5, 8

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 4

[15] Chi-Jui Ho, Chin-Cheng Chan, and Homer H. Chen. Af-net: A convolutional neural network approach to phase detection autofocus. *IEEE TIP*, 2020. 2

[16] Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *ICML*, 2017. 2

[17] Jui-Ting Huang, Chun-Hung Shen, See-May Phoong, and Homer Chen. Robust measure of image focus in the wavelet domain. *IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, 2005. 2

[18] Nasser Kehtarnavaz and H-J. Oh. Development and real-time implementation of a rule-based auto-focus algorithm. *Real-Time Imaging*, 2003. 2

[19] Masahiro Kobayashi, Michiko Johnson, Yoichi Wada, Hiromasa Tsuboi, Hideaki Takada, Kenji Togo, Takafumi Kishi, Hidekazu Takahashi, Takeshi Ichikawa, and Shunsuke Inoue. A low noise and high sensitivity image sensor with imaging and phase-difference detection af in all pixels. *ITE Transactions on Media Technology and Applications*, 2016. 2

[20] Matej Kristan, Janez Pers, Matej Perse, and Stanislav Kovacic. A bayes-spectral-entropy-based measure of camera focus using a discrete cosine transform. *Pattern Recognition Letters*, 2006. 2

[21] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017. 2

[22] Sang Yong Lee, Jae Tack Yoo, and Soo-Won Kim. Reduced energy-ratio measure for robust autofocusing in digital camera. *Signal Processing Letters*, 2009. 2

[23] Sang-Yong Lee, Yogendera Kumar, Ji-Man Cho, Sang-Won Lee, and Soo-Won Kim. Enhanced autofocus algorithm using robust focus measure and fuzzy reasoning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2008. 2

[24] Amir Pouran Ben Veyseh, Ning Xu, Quan Tran, Varun Manjunatha, Franck Dernoncourt, and Thien Nguyen. Transfer learning and prediction consistency for detecting offensive spans of text. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *ACL*, 2022. 3

[25] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992. 5

[26] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. In *ICLR*, 2018. 2

[27] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *NeurIPS*, 2016. 2

[28] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 5

[29] Przemyslaw Sliwinski and Pawel Wachel. A simple model for on-sensor phase-detection autofocusing algorithm. *Journal of Computational Chemistry*, 2013. 2

[30] Dong-Chen Tsai and Homer H Chen. Smooth control of continuous autofocus. In *ICIP*, 2012. 2

[31] Chengyu Wang, Qian Huang, Ming Cheng, Zhan Ma, and David J Brady. Deep learning for camera autofocus. *IEEE Transactions on Computational Imaging*, 2021. 2

[32] Hui Xie, Weibin Rong, and Lining Sun. Wavelet-based focus measure and 3-d surface reconstruction method for microscopy images. In *IROS*, 2006. 2

[33] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *NeurIPS*, 2020. 2

[34] Xin Xu, Yinglin Wang, Jinshan Tang, Xiaolong Zhang, and Xiaoming Liu. Robust automatic focus algorithm for low contrast images using a new contrast measure. *Sensors*, 2011. 2

[35] Xin Xu, Xiaolong Zhang, Haidong Fu, Li Chen, Hong Zhang, and Xiaowei Fu. Robust passive autofocus system

for mobile phone camera applications. *Computers & Electrical Engineering*, 2014. 2

[36] Ge Yang and Bradley J Nelson. Wavelet-based autofocusing and unsupervised segmentation of microscopic images. In *IROS*, 2003. 2

[37] Yi Yao, Besma Abidi, Narjes Doggaz, and Mongi Abidi. Evaluation of sharpness measures and search algorithms for the auto-focusing of high magnification images. *Pattern Recognition Letters*, 2006. 2

[38] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *ICCV*, 2019. 2

[39] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 2