LEARNING ADJUSTABLE IMAGE RESCALING WITH JOINT OPTIMIZATION OF PERCEPTION AND DISTORTION

Zhihong Pan

Baidu Research (USA), Sunnyvale, CA, 94089, USA

ABSTRACT

The performance of image super-resolution (SR) have been greatly advanced by deep learning techniques recently. Most models are only optimized for the ill-posed upscaling task while assuming a predefined downscaling kernel for low-resolution (LR) inputs. Additionally, there exists a conflict between the objective and perceptual qualities of upscaled outputs for optimizing these models. To achieve an effective trade-off between these two qualities, the current methods are either inflexible as the model is optimized for a fixed trade-off, or inefficient as it needs to interpolate weights or images from two separately trained models. Based on the invertible rescaling net (IRN) which learns image downscaling and upscaling together, we propose a joint optimization method to train just one model that could achieve adjustable trade-off between perception and distortion for upscaling at inference time. Additionally, it's shown in experiments that this jointly optimized model could produce results with better accuracy while maintaining high perceptual quality compared to one optimized for perceptual quality only.

Index Terms— Super-Resolution, Image Rescaling

1. INTRODUCTION

Image super-resolution (SR), the process to recover high-resolution (HR) images from low-resolution (LR) inputs, is an ill-posed challenge as there exist multiple possible HR images resulting in the same downscaled LR image. Lately, the powerful deep learning techniques have led to developments of many single image super resolution models [1, 2] reporting impressive performances. To achieve best reconstruction accuracy, these methods aims to minimize the mean squared error (MSE) or L1 loss between the restored and the ground-truth (GT) images. As this line of work tries to accommodate the mapping between one LR and multiple HR outputs, it often generates blurred output which lacks sharp details as in real HR images. While it is ideal to restore a SR image which is both accurate and photo-realistic, there is a trade-off between the ability to achieve low MSE and high perceptual quality as pointed out in [3]. So for methods aim to improve the perceptual image quality using adversarial training [4, 5], the outputs are sharper but are subjected to lower accuracy when compared to GT references.

To tackle this challenge, both SRGAN-MSE [4] and ENet [5] try to train the model with a mixture of MSE and adversarial losses, but they lead to unstable training due to the natural conflict between two losses which causes undesirable artifacts. These methods are also inflexible as the mixture of losses is part of hyperparameters determined before training. Alternatively, other methods try to achieve the best trade-off using two models. ESRGAN [6] proposed to train two separate networks which enhance the objective and perceptual quality respectively and combine them using weights interpolation. Using image style transfer, Deng [7] proposed to fuse images from two models for better results. Later, Deng *et al*. [8] further proposed to implement the style transfer in the wavelet domain so the objective and perceptual qualities can be preserved in different wavelet subbands for improved performance. However, the above two-model methods are not computationally efficient for training, as well as for inference when additional image interpolation or fusion is needed.

Another drawback of many image SR models originates from its dependence on training with LR-HR image pairs with the LR inputs are synthesized from a predefined downscaling kernel. Furthermore, the models are trained for upscaling reconstruction only without taking the image downscaling method into consideration together. To take advantage of the potential mutual beneficiary reinforcement between downscaling and the inverse upscaling, Kim *et al*. [9] proposed an auto-encoder framework to jointly train image downscaling and upscaling together. Similarly, Sun *et al*. [10] proposed a new content adaptive-resampler based image downscaling method, which can be jointly trained with any existing differentiable upscaling (SR) models. More recently, Xiao *et al*. [11] proposed a invertible rescaling net (IRN) that has set the state-of-the-art (SOTA) for learning based image rescaling. Based on the invertible neural network (INN) [12], IRN learns to convert HR input to LR output and an auxiliary latent variable z. By mapping z to a case-agnostic normal distribution during training, inverse image upscaling is implemented by randomly sampling z from the normal distribution without need of the case specific $\hat{z}$. Using different losses at training, IRN can be optimized for either objective or perceptual quality. Interpolation or fusion of two images from two separately optimized models can be used to achieve a good

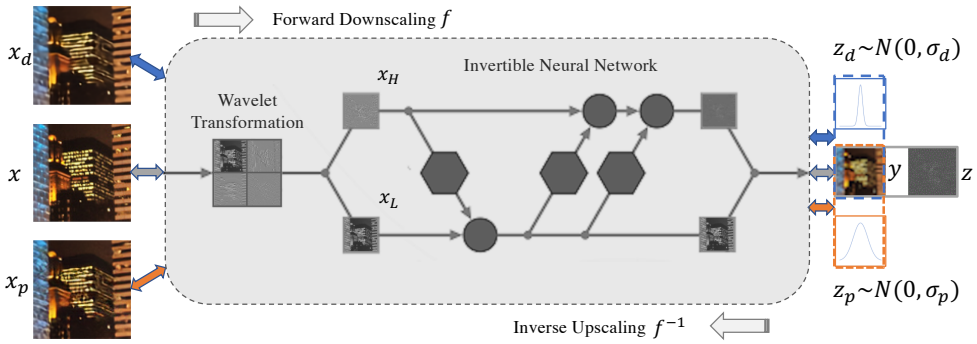

Fig. 1: Network architecture for invertible image rescaling with adjustable trade-off between perception and distortion.

trade-off between perception and distortion.

To overcome this inefficiency, a joint optimization method is proposed here to train one joint model that can be used to achieve adjustable trade-off between perception and distortion at inference. Based on IRN, a mixture of reconstruction, perceptual and adversarial losses is used for training and the upscaling related losses are conditioned on the randomly sampled auxiliary latent variable z during joint optimization. At inference, upscaled outputs can be adjusted by modulating the random sampling of z. The key highlights of the proposed method include:

- As far as we know, it is the first solution to train one image rescaling model with joint optimization for both objective and perceptual qualities.
- The trained model is capable of generating upscaled images, in one inference, with adjustable trade-off between the two qualities using latent feature modulation.
- The trained model is able to produce upscaled image with less distortion comparing to the one trained solely for perception, without suffering in perceptual quality.

2. PROPOSED METHOD

2.1. Problem Formulation

As shown in Fig. 1, the network architecture of previous IRN work [11] is utilised for our proposed adjustable image rescaling method. First, the input HR image x is split to low-frequency component x_L and high-frequency component x_H using an $\times 2$ wavelet transformation. The split components are then transformed into a downscaled LR image y and an auxiliary latent variable z using an invertible neural network. This forward downscaling process is described as $(y, z) = f(x)$. As both steps are invertible transformations, the inverse upscaling process $x = f^{-1}(y, z)$ is determined once f is known. Note that multiple units of such blocks are cascaded to achieve larger scaling factors.

Due the ill-posed nature of upscaling, there are multiple versions of HR outputs corresponding to variations in z. As showing in Fig.1, for certain z_d, the upscaled $x_d = f^{-1}(y, z_d)$ is similar to x with little distortion, but has blurry effects and lower perceptual quality. While for x_p as generated from (y, z_p), it has sharper details for better perception but suffers loss in accuracy. In previous IRN work, it forces z to be a case-agnostic random variable thus the inverse upscaling process can reconstruct an HR image $\hat{x}$ to either best objective quality like x_d or perceptual quality like x_p using different mixtures of losses as shown below

$$L = \lambda_1 L_r + \lambda_2 L_g + \lambda_3 L_d + \lambda_4 L_p. \quad (1)$$

Here L_r is the $L1$ reconstruction loss for upscaled HR output and L_g is the $L2$ guidance loss for downscaled LR output. For L_d, it is either implemented as distribution regulation of of latent variable z to help train a model (IRN) with minimal distortion, or as an adversarial loss from a co-trained discriminator to train a model (IRN+) generating photo-realistic HR images. The perceptual loss L_p is only used for IRN+ when the goal is best perceptual quality. While the proposed losses work well for either optimal objective or perceptual qualities, it is not an efficient solution. Not only it requires training of two separate models, the random sampling of z from normal distribution $N(0, 1)$ does little in increasing diversity of upscaled outputs. In the case of IRN, differences in PSNR results from different z samples are less than 0.02. Here we propose a joint optimization method to train just one model that can generate multiple HR outputs, one with high accuracy, or high perception quality, or an adjustable trade-off between the two by sampling z differently.

2.2. Joint Optimization

For learned image rescaling problems, it is required that only the quantized image of downscaled y can be used as known input for inverse upscaling. For IRN, as latent variable z is also needed as input for upscaling, a case-agnostic random variable is used at inference. To achieve our goal of flexible trade-off between perception and distortion with one model, additional information from z must be utilized. For randomly sampled z, there are only two associated parameters μ and σ. On the other hand, it is known in information

theory [13] that the differential entropy of a normal distribution is $\ln(\sigma\sqrt{2\pi e})$, which depends on the standard deviation σ solely. In other words, The entropy of normally distributed z is higher when z has higher variance, a larger σ. Lastly, it is known that a restored a HR image x_d with little distortion is more blurry while x_p, which is optimized for better perception, has sharper details. In other words, x_p has higher entropy than x_d.

Inspired by these observations, a joint optimization method is proposed to train an IRN model using the following loss

$$L = (1-\sigma)\lambda_1 L_r + \lambda_2 L_g + \sigma(\lambda_3 L_d + \lambda_4 L_p). \quad (2)$$

Here the four individual losses, including associated weights λ_i, are the same ones used for IRN+ training. The forward downscaling process is the same as in IRN for our method. For upscaling, z is randomly sampled from $N(0,\sigma)$. Empirically, we limit σ in the range of (0, 1) during training. Thus for smaller σ, latent variable z has lower entropy, the model is biased towards minimizing reconstruction loss L_r, which leads to restored HR output with less distortion and relatively lower entropy. When σ is larger and z has higher entropy, the model is biased towards perception related losses L_d and l_p.

After the joint optimization completed, based on desired trade-off between perception and distortion, the upscaled image can be flexibly generated as $x_\sigma = f^{-1}(y, z_\sigma)$ when z_σ is randomly sampled from $N(0,\sigma)$. Here σ could be set as 0 for the least distortion, or as 1 for best perception, or any value in between according to the desired trade-off. This is done at inference time of upscaling and only one model is needed.

3. EXPERIMENTS

For fair comparison, we use the same training strategy and settings as in previous IRN work [11], including training dataset DIV2K [14] and hyperparamers like batch size. The proposed joint optimization was applied to the pre-trained IRN+ model in [11]. There are a total 200,000 iterations in training, from an initial learning rate of 1×10^{-4} which decays by half after every 40,000 iterations. The trained model is denoted as IRN$_\sigma$ here for convenience.

For quantitative assessment of objective qualities, we use the peak noise-signal ratio (PSNR) and SSIM [15] on the Y channel in the YCbCr color space. For perception, the learned perceptual image patch similarity (LPIPS) [16] metric is selected for its recent successful application to use as a metric of perceptual quality when the GT reference is known.

3.1. Optimization Strategy

As elaborated in Section 2.2, the joint optimization process is implemented by rotating different σ for $z = N(0,\sigma)$ at training and using conditioned losses that rely on σ as a parameter. As shown in Fig. 2, three different sampling strategies are compared using perception-distortion trade-off curve of DIV2K validation set in $\times 4$, which is plotted by connecting points in PSNR-LPIPS coordinates when setting σ from 0 to 1 with 0.1 increments during testing. As shown in grey, when σ is drawn from an uniform distribution between 0 and 1 at training, it has the best trade-off between objective and perceptual qualities comparing to two other sampling strategies as it is closer to the lower-left corner overall. However, this curve collapses to a very small range, and the individual peak value of PSNR and LPIPS is significantly worse than the other two. For the 2-point sampling strategy, where σ is either 0 or 1, it can generate the highest PSNR and lowest LPIPS individually but the trade-off in the middle is not satisfying. The 3-point (0, 0.5 and 1) sampling is the best overall, capable of equivalent individual peak performance while enjoying greatly improved trade-off comparing to 2-point sampling. IRN$_\sigma$ used in all following assessments is trained with 3-point sampling of σ unless specified otherwise.

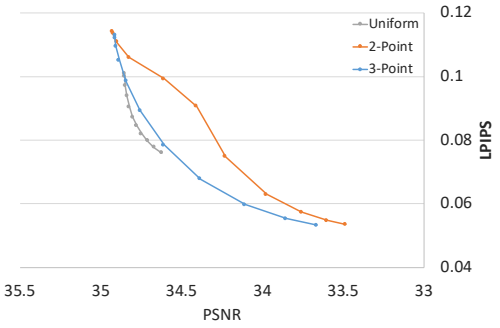

Fig. 2: Perception-distortion trade-off comparison for different σ sampling strategies during training.

3.2. Quantitative and Qualitative Results

The perception-distortion trade-off curves are plotted for three processes in Fig. 4. For the first one, RCAN [2] is chosen as the SR baseline for minimum distortion while ESRGAN [6] is chosen for best perception. The trade-off curve is generated by interpolating images from two models. The IRN-IRN+ one is generated similarly from distortion optimal IRN and perception optimal IRN+. As both IRN and IRN+ are jointly optimized for bidirectional downscaling and upscaling, it outperforms the RCAN-ESRGAN SR baseline, which are optimized for upscaling only, by a remarkably large margin for both objective and perceptual metrics. Note that both processes use interpolation from two individually trained models. In contrast, only one jointly-optimized model is need to generate the trade-off curve of IRN$_\sigma$ by setting different σ values to generate z_σ during upscaling. Comparing to IRN-IRN+, the curve is shifted to the lower-left in general, meaning improved performance in both distortion and perception. While it has a slightly lower peak PSNR value, its peak LPIPS value is actually better than IRN+ while achieving significantly less distortion (higher PSNR). This

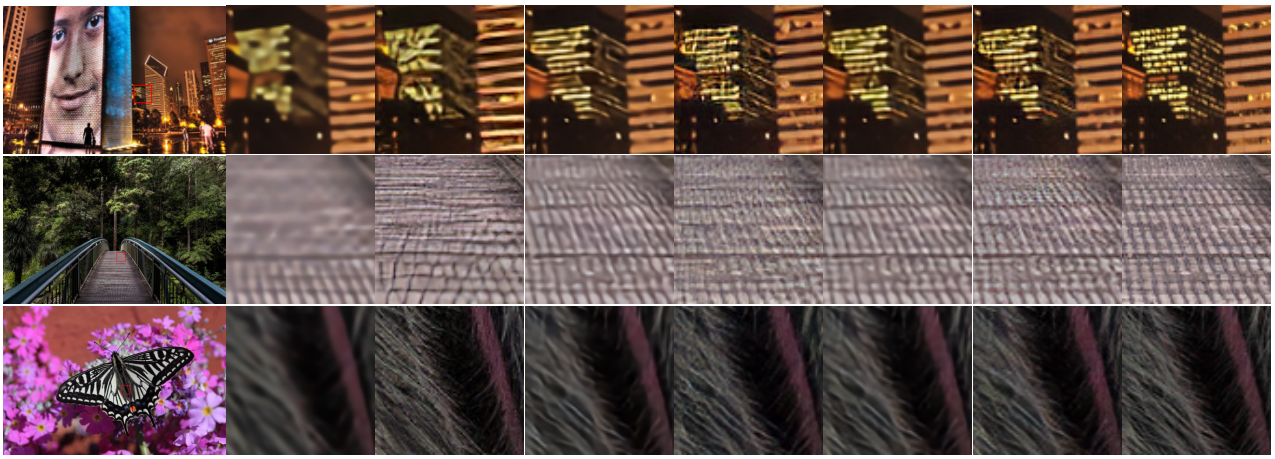

Fig. 3: Visual examples from Urban100 test set and DIV2K validation set (Best viewed in online version with zoom-in).

Table 1: Comparison of objective and perceptual qualities for upscaled $\times 4$ images of 5 datasets, with the best two results highlighted in red and blue. IRN$_{\sigma:0}$ and IRN$_{\sigma:1}$ refer to results from IRN$_\sigma$ while fixing σ at 0 and 1 during inference respectively.

Method	Set5 [17]		Set14 [18]		BSD100 [19]		Urban100 [20]		DIV2K [21]	
	LPIPS↓	PSNR/SSIM↑	LPIPS↓	PSNR/SSIM↑	LPIPS↓	PSNR/SSIM↑	LPIPS↓	PSNR/SSIM↑	LPIPS↓	PSNR/SSIM↑
RCAN [2]	0.1695	32.64/0.899	0.2740	28.85/0.788	0.3589	27.74/0.742	0.1967	26.75/0.806	0.2547	30.72/0.844
ESRGAN [6]	0.0726	30.46/0.851	0.1323	26.28/0.697	0.1630	25.29/0.649	0.1239	24.35/0.732	0.1150	28.17/0.775
IRN [11]	0.0782	36.19/0.944	0.1237	32.67/0.901	0.1654	31.63/0.881	0.0836	31.40/0.915	0.1174	35.07/0.931
IRN+ [11]	0.0312	33.63/0.914	0.0668	29.97/0.843	0.0749	28.93/0.818	0.0550	28.24/0.867	0.0541	32.24/0.891
IRN$_{\sigma:0}$	0.0726	36.08/0.943	0.1178	32.54/0.898	0.1626	31.48/0.879	0.0798	31.20/0.913	0.1130	34.91/0.929
IRN$_{\sigma:1}$	0.0336	34.79/0.924	0.0664	31.27/0.870	0.0778	30.17/0.848	0.0487	30.19/0.896	0.0534	33.66/0.910

advantage of increased PSNR for outputs of high perceptual qualities is consistent throughout different datasets as shown in Table 1. Results from IRN$_{\sigma:1}$ have the lowest LPIPS for 3 out of 5 datasets and only falling behind IRN+ slightly for the other two, while leading IRN+ in PSNR by 1.16 or more for all 5 datasets. For objective metrics PSNR and SSIM, IRN$_{\sigma:0}$ is slightly behind IRN as the second best, trailing by 0.11 to 0.20 in PSNR for all 5 test sets.

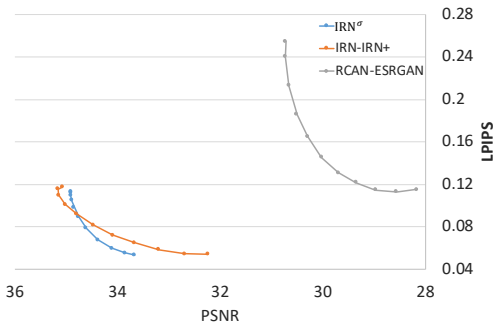

Fig. 4: Perception-distortion trade-off comparison of different image SR and rescaling models (DIV2K $\times 4$).

For qualitative assessment, visual examples from Urban100 and DIV2K are compared in Fig. 3. While ESRGAN produces much sharper image than RCAN, it is subject to compromised accuracy, like distorted line patterns in the first two examples. The other four are all image rescaling models and their restoration qualities exceed both RCAN and ESRGAN greatly. For IRN and IRN$_{\sigma:0}$, the details are relatively more blurry but no big difference between the two. While both IRN and IRN$_{\sigma:0}$ generates images with equivalent perceptual quality, IRN+ tends to introduce more artifacts, like adding high frequency details in the middle example, and causing curvatures in straight lines of the first one.

4. CONCLUSIONS

A simple and effective joint optimization method is proposed to tackle the challenge in learned image rescaling: the conflict between maximization of objective and perceptual qualities. Using existing IRN backbone, the proposed method introduces losses conditioned on the random sampling of latent variable z. The model is trained to minimize distortion loss more when z has a lower entropy (smaller σ in $N(0, \sigma)$) and optimized for lower perception loss when σ is larger. Using a 3-point sampling of σ at training, one optimized IRN$_\sigma$ can generate multiple upscaled images with adjustable trade-off between perception and distortion from one input, which previously was only possible with two separate models.

5. REFERENCES

[1] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*. Springer, 2014, pp. 184–199.

[2] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 286–301.

[3] Yochai Blau and Tomer Michaeli, "The perception-distortion tradeoff," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6228–6237.

[4] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[5] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch, "EnhanceNet: Single image super-resolution through automated texture synthesis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4491–4500.

[6] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 63–79.

[7] Xin Deng, "Enhancing image quality via style transfer for single image super-resolution," *IEEE Signal Processing Letters*, vol. 25, no. 4, pp. 571–575, 2018.

[8] Xin Deng, Ren Yang, Mai Xu, and Pier Luigi Dragotti, "Wavelet domain style transfer for an effective perception-distortion tradeoff in single image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3076–3085.

[9] Heewon Kim, Myungsub Choi, Bee Lim, and Kyoung Mu Lee, "Task-aware image downscaling," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 399–414.

[10] Wanjie Sun and Zhenzhong Chen, "Learned image downscaling for upscaling using content adaptive resampler," *IEEE Transactions on Image Processing*, vol. 29, pp. 4027–4040, 2020.

[11] Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaolong Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu, "Invertible image rescaling," in *European Conference on Computer Vision*. Springer, 2020, pp. 126–144.

[12] Lynton Ardizzone, Jakob Kruse, Sebastian Wirkert, Daniel Rahner, Eric W Pellegrini, Ralf S Klessen, Lena Maier-Hein, Carsten Rother, and Ullrich Köthe, "Analyzing inverse problems with invertible neural networks," *arXiv preprint arXiv:1808.04730*, 2018.

[13] Thomas M Cover, *Elements of information theory*, John Wiley & Sons, 1999.

[14] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani, "Blind super-resolution kernel estimation using an internal-gan," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[15] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[16] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[17] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie line Alberi Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proceedings of the British Machine Vision Conference*. 2012, pp. 135.1–135.10, BMVA Press.

[18] Roman Zeyde, Michael Elad, and Matan Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*. Springer, 2010, pp. 711–730.

[19] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*. IEEE, 2001, vol. 2, pp. 416–423.

[20] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5197–5206.

[21] Eirikur Agustsson and Radu Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 126–135.