ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446201

SR-VFA: ACCURATE SELF-REFINED FACE ALIGNMENT IN VIDEOS

Sipeng Yang *Hongyu Huang* *Qingchuan Zhu* *Xiaogang Jin*

State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, China
{*12121024, 3210105703, 22251012*}*@zju.edu.cn, jin@cad.zju.edu.cn*

ABSTRACT

Face alignment is a critical and difficult task for many facial analysis applications. Existing VFA methods frequently ignore the consistency of facial geometries and textures across video sequences, limiting their ability to handle accurate and stable face alignment. This paper describes a robust and highly accurate 3D Morphable Model (3DMM)-based VFA approach that employs a novel texture generation method and a self-refined face alignment procedure. Our method iteratively fine-tunes facial geometries, textures, and poses by using a differentiable rendering technique and a self-refined optimization method. Experiment results show that our method outperforms existing state-of-the-art methods in terms of both accuracy and temporal stability. Visual results and source code are available at: `https://pawindergit.github.io/SR-VFA/`

Index Terms— Video-based face alignment, 3D dense face alignment, self-refining model

1. INTRODUCTION

Face alignment (FA) involves accurately identifying specific facial landmarks, such as the corners of the eyes and mouth, within facial images or video frames. This fundamental step is crucial for a range of facial analysis tasks, including face action unit detection [1], expression and micro-expression recognition [2], and face reconstruction [3]. Its importance in ensuring the stability and precision in these applications has driven the pursuit for more accurate and robust FA solutions.

Depending on the type of data addressed, FA approaches can be broadly divided into two categories: those designed for single images and those tailored for videos. Over the past few decades, single-image face alignment has received considerable attention and achieved impressive results in both speed and accuracy. Dominant methods in this area include cascaded regression [4], deep learning heatmap prediction[5], and 3D face model fitting [6]. In contrast, video-based face alignment (VFA) research [7], which aims to track facial landmarks in successive video frames, has been relatively less explored. An additional challenge unique to VFA primarily lies in the heightened requirement for consistency in the predicted facial landmarks across adjacent frames.

Exploiting the temporal continuity of faces in video sequences [8] is a straightforward and prevalent VFA approach for robust facial landmark detection. Specifically, recurrence regression-based [9], optical flow-based [10] , and cycle-consistency-based [11] methods propose reusing the outputs or intermediate features from one frame to facilitate precision landmarks prediction in subsequent frames. But these methods commonly neglect the consistency of facial geometries within video sequences, thus limiting their effectiveness in addressing faces with occlusions or extensive motion. To address this issue, other VFA approaches based on 3DMM utilize reconstructed facial geometries to achieve video face alignment. These methods typically impose constraints on the identity parameters of facial models within a single video or multiple perspectives, subsequently predicting or optimizing facial expressions, positions, and orientations for face alignment. However, despite the advantages of identity consistency provided by 3DMM, these methods often struggle to accurately align 3DMM textures with real-world facial features, which undermines the precision and reliability of thes VFA approaches.

In this paper, we present a novel 3DMM-based VFA method aiming to achieve robust and highly accurate face alignment in video sequences. Our key innovation involves sampling and generating realistic facial textures in video frames, complemented by the iterative refinement of facial alignment through a self-refined procedure. We employ the Basel Face Model (BFM) [12] as the foundational framework for face reconstruction. Given that the BFM supports only low-frequency textures, its ability to accurately represent diverse facial features is restricted. This limitation leads to discrepancies between the 3DMM models generated and the actual faces in video frames, thereby compromising the precision of aligned facial positions and orientations. To address this issue, we introduce a sampling-based approach to acquire more realistic facial textures. Subsequently, the facial positions and orientations, along with the facial textures, are iteratively updated using the differentiable rendering technique. Our method leads to robust and highly accurate face alignment across video sequences. The main contributions of this work include (1) a novel sampling-based approach for acquiring realistic facial textures and (2) a new self-refined procedure for enhanced video-based face alignment.

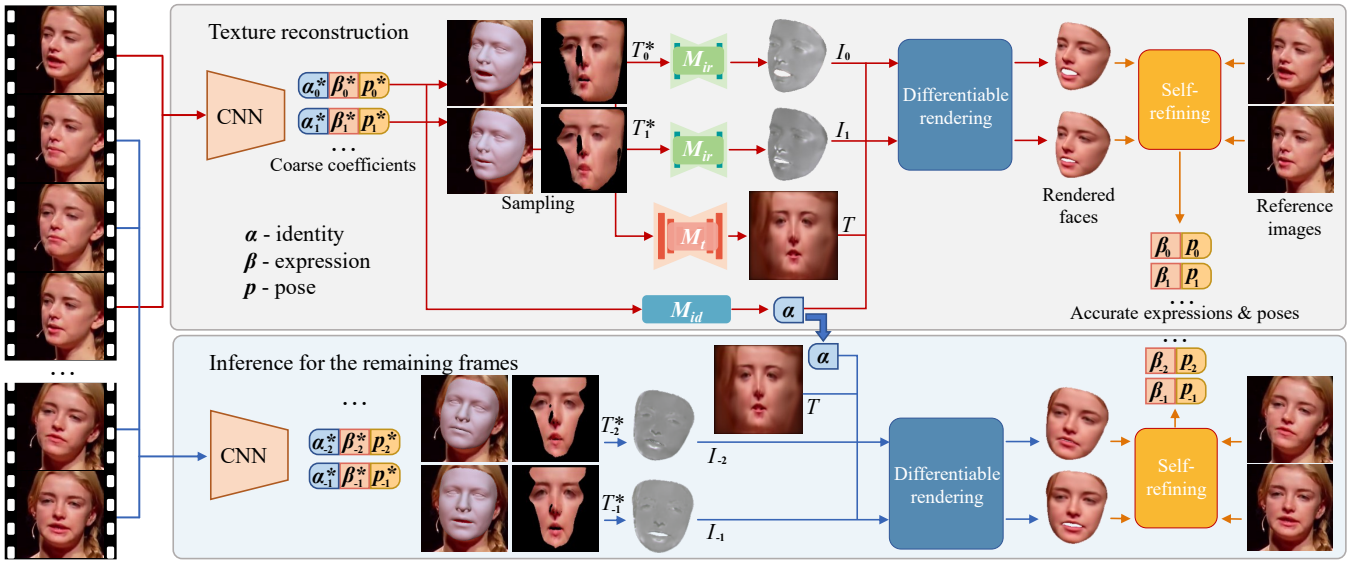

Fig. 1. Overview of the proposed VFA method. Our approach comprises two phases: texture reconstruction and frame inference. Details about the network structures are available on the project page.

2. METHODOLOGY

Traditional 3DMM methods often yield facial reconstructions that lack realism. Despite recent studies [13] incorporating refined Bidirectional Reflectance Distribution Function (BRDF) to enhance the rendered results, noticeable discrepancies persist between rendered and real facial appearances. To address this issue while optimizing reconstruction speed and alignment accuracy, our method introduces a simplified face model with coefficients and features, including identity α, expression β, and pose p from 3DMM, along with texture T sampled from video frames and texture-free irradiance I. This adjustment enables realistic 3D face modeling, thereby contributing to following precise face alignment.

Figure 1 illustrates the comprehensive pipeline of our face alignment method. Given a video sequence containing faces of a consistent identity, we partition the sequence into two sets. The first set consists of a smaller subset of frames that are evenly sampled to construct a realistic facial texture. The second set comprises the remaining frames, which exploit this pre-constructed texture to achieve high-precision face alignment, obviating the need for additional texture modeling.

In the texture reconstruction phase, an initial coarse facial geometry is obtained for each selected frame using the BFM [12]. These facial geometries serve as the foundational basis for sampling facial appearances, which are subsequently refined into a photorealistic texture via Gated Convolutional Neural Networks (CNNs). In parallel, these sampled facial appearances are utilized to compute texture-free irradiances. Finally, through differentiable rendering and self-refining optimization, the 3D facial geometries, photorealistic texture, and texture-free irradiances are synthesized to achieve precise face alignment in each frame. This method produces consistent and highly accurate face alignment across the video sequence. The generated textures are reused for subsequent frames to facilitate rapid face alignment, eliminating the need for additional texture reconstruction. The texture reconstruction phase remains consistent with the alignment process.

2.1. Texture reconstruction

Coarse face coefficients. We start with a CNN based model to estimate coarse coefficients for 3DMMs of selected frames. These coefficients contain basic information about facial identity α_i^*, expression β_i^*, and pose p_i^* of frame i. To enhance both efficiency and robustness of the subsequent face alignment model, we pretrained the coefficient-estimation model on a dataset of in-the-wild face images following [14].

Video-consistent identity and texture. Theoretically, variations in facial expressions, poses, and lighting naturally occur across different frames of a video sequence, while identity coefficient and texture should exhibit temporal consistency. Motivated by this, a uniformly selected subset of frames is used for the computation of a temporally consistent texture image and identity coefficient. For the texture generation, coarse 3D facial models reconstructed from these selected frames serve as the foundation for texture sampling. It can be observed that the sampled texture maps $\{T_0^*, T_1^*, \ldots\}$ are largely similar and may miss occluded portions. Therefore, a model M_t equipped with gated convolutional layers is deployed to generate a unified facial texture T specific to the video sequence. As for the identity component, the aim is to amalgamate a set of coefficients $\{\alpha_0^*, \alpha_1^*, \ldots\}$ from different frames into a video-consistent identity coefficient α. Given these coeffi-

cients are inherently similar but may contain minor fluctuations, it is logically concluded that a transformer model M_{id} is apt for this specific task. Note that positional encoding is excluded from the transformer model due to the weak temporal correlations among the selected frames.

Self-refined expression and pose. After acquiring the texture image T and identity coefficient α for a given video sequence, we construct the 3D face models F_i for each frame i as:

$$F_i = \begin{cases} \text{geometry:} & \text{BFM}(\alpha, \beta_i, p_i), \\ \text{shaded texture:} & T * I_i, \end{cases}$$

where the "geometry" part is derived from the BFM method [12]. The "shaded texture" combines the texture T with the texture-free irradiance I_i. The latter serves as an indicator of texture brightness and is constrained to the range $[0, 10]$. To generate I_i, we employ a lightweight CNN model M_{ir} equipped with gated convolutional layers to convert the sampled images to texture-free irradiances. With the necessary components prepared, the 3D face model F_i is rendered into a 2D image R_i and aligned with its corresponding video frame P_i. We formulate the alignment task as an optimization problem involving the expression $\beta_i = \beta_i^* + \Delta\beta_i$ and pose $p_i = p_i^* + \Delta p_i$, as well as the trainable models M_t, M_{id}, and M_{ir}. Differentiable rendering techniques are used to enable gradient backpropagation. The objective function $f(i)$ comprises a structural similarity (SSIM) term [15], augmented by three L_2 regularization terms for faster convergence: $f(i) = 1 - \text{SSIM}(R_i, P_i) + \lambda_1\|\Delta\beta_i\|_2 + \lambda_2\|\Delta p_i\|_2 + \lambda_3\|I_i - 1\|_2$, where weights λ_1, λ_2, and λ_3 are empirically set to 0.1, 0.1 and 0.5, respectively. The optimization is conducted using the Adam optimizer [16], with a learning rate decaying from 1e-4 to 1e-6 over 1,000 epochs.The batch size, set to 128, also corresponds to the number of frames selected for texture generation phase. Upon optimization, we obtain the self-refined facial texture T, identity coefficient α of the given video, and precise alignment for pose p and expression β of each frame. Fig. 2 provides an illustration of the intermediate results.

2.2. Inference for the remaining frames

A given video can contain a large number of facial frames. To enhance computational efficiency, we select only a small subset of these frames for the generation of the temporally consistent texture image T and the identity coefficient α. The remaining frames are then efficiently aligned by leveraging the precomputed T and α. As shown in the lower portion of Fig. 1, we employ a self-refining method same to the one described in Sec. 2.1 to accurately estimate pose and expression coefficients. During this stage, optimization is only focused on the expression β_i, the pose p_i, and the texture-free irradiance I_i. All other computational settings, including the choice of objective function and optimizer, are consistent with those outlined in Sec. 2.1.

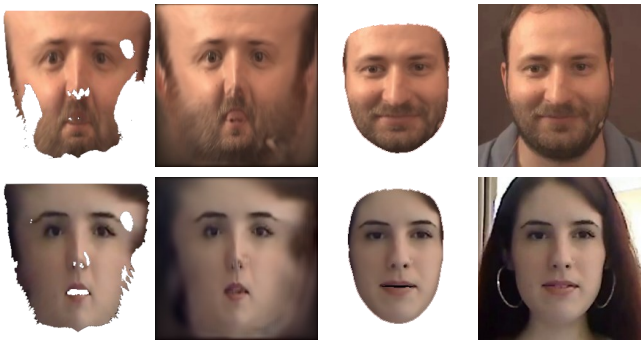

Fig. 2. Illustration of intermediate results. From left to right, the figure displays the sampled textures, the reconstructed textures, the final 3D face models, and reference images.

2.3. Inference time

We have optimized our approach and designed lightweight models, ensuring manageable inference times. The first phase, optimizing facial texture, takes about *1 minute for one person* on an RTX 3090 GPU. The second phase, focusing on facial pose and expression, is faster, processing more than *180 frames per minute* on the mentioned hardware.

3. EXPERIMENTS

3.1. Dataset and evaluation metrics

We use the 300-VW dataset [21], consisting of 114 videos totaling 218,595 frames, to evaluate our VFA method and compare it with state-of-the-art (SOTA) approaches. Of these, 64 videos serve as the test set, which are categorized into three difficulty levels (A, B, and C), with category C being the most challenging as its inclusion of low-resolution and poor-quality faces. The test videos cover a diverse set of facial expressions and poses, offering a comprehensive evaluation foundation.

To evaluate the face alignment results of different methods, we employ two widely-used metrics in this field: the normalized mean error (NME) and the normalized mean flicker (NMF) [9]. The NME is a standard metric that quantifies the mean discrepancy between the predicted and ground-truth (GT) landmarks. Given N frames, each containing L GT landmarks, let $m_{i,l}$ and $\hat{m}_{i,l}$ denote the GT and predicted 2D coordinates for the i-th frame and l-th landmark, respectively. The residual vector is defined as $\mathbf{r}_{i,l} = m_{i,l} - \hat{m}_{i,l}$. The NME is then computed as: $NME = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{L}\sum_{l=1}^{L}\frac{\|\mathbf{r}_{i,l}\|}{d_0}\right)$, where d_0 is the inter-ocular distance to normalize the error with respect to human perception. While the NMF metric is designed to measure the temporal coherence of landmark positions across frames and is defined as: $NMF = \sqrt{\frac{1}{N}\sum_{i=2}^{N}\left(\sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(\frac{\|\mathbf{r}_{i,l}-\mathbf{r}_{i-1,l}\|}{d_1}\right)^2}\right)^2}$, where d_1^2 is the

Methods	Metrics	FAN [17] ICCV 17'	SBR [8] CVPR 18'	3DDFAv2 [7] ECCV 20'	MGC [18] ECCV 20'	DECA [3] ToG 21'	PIP [19] IJCV 21'	SLPT [20] CVPR 22'	LDEQ [9] CVPR 23'	Ours
Category A	NME	4.210	4.175	4.968	3.945	3.882	3.451	3.454	<u>3.388</u>	**3.341**
	NMF	254.42	154.28	174.07	<u>132.86</u>	133.66	139.68	146.22	138.73	**131.97**
Category B	NME	3.988	3.755	4.667	4.054	4.219	3.697	3.640	**3.472**	<u>3.546</u>
	NMF	295.55	192.04	212.73	274.79	180.11	169.83	171.38	<u>161.67</u>	**149.86**
Category C	NME	7.416	6.267	6.891	6.295	6.648	5.951	**5.352**	6.047	<u>5.564</u>
	NMF	414.74	240.52	272.08	392.51	298.19	290.14	337.53	<u>270.62</u>	**246.95**

Table 1. Comparison of FA methods across three categories. NME and NMF are used to assess the accuracy and temporal stability of facial landmarks, respectively. Best results are highlighted in bold, and second-best results are underlined.

Optimizing	Identity	Texture	NME	NMF	SSIM
✗	✗	✗	5.432	276.51	0.8610
✓	✗	✗	4.017	164.44	0.8849
✓	✓	✗	3.882	161.19	0.8943
✓	✓	✓	**3.763**	**157.80**	**0.9512**

Table 2. Ablation study assessing contributions of proposed method components using NME, NMF, and SSIM metrics.

face area and serves for normalization. The root mean square is applied to penalize abrupt changes more effectively, thus better capturing the human perception of flicker.

3.2. Comparison

We compare our proposed model against SOTA methods across the 3 categories on the 300-VW dataset. Our baselines include the image face alignment methods [17, 19, 20], optical flow-based [8] and 3DMM-based [7, 18, 3], and heat map recurrence regression-based [9] VFA methods. The performance of these methods is evaluated in terms of both accuracy and temporal stability of the extracted facial landmarks, using the metrics NME and NMF.

Tab. 1 shows the performance of various face alignment methods across test sets of differing complexity. In Category A, which corresponds to the easiest scenarios, our method and LDEQ [9] achieve the lowest NME scores which indicates higher accuracy in the positions of detected facial landmarks. Meanwhile, our method and MGC [18] attain the lowest NMF scores, which indicating minimal temporal fluctuations and greater stability in landmark predictions. For Category B, our method and LDEQ [9] exhibit strong performance across both the NME and NMF metrics. In the most challenging Category C, our method outperforms other approaches by achieving both a lower NME score and markedly better NMF scores. Overall, our VFA method exhibits performance that is comparable or superior to existing SOTA methods across all categories. Especially for the temporal stability metric NMF, our approach consistently outperforms existing methods. These results validate the accuracy and robustness of our method in addressing the video-based face alignment tasks.

3.3. Ablation Study

An ablation study is conducted to assess the contributions of individual components in the proposed method. We employ the coefficient-estimation CNN model as our initial baseline for comparison. Then, we incrementally augment this baseline by introducing the following enhancements: (1) optimizing for facial expressions and pose only; (2) incorporating temporally consistent identity coefficients; and (3) adding generated texture maps for further refinement. For the evaluation, a subset of 30 videos is randomly selected from the 300-VW dataset. Besides the NME and NMF metrics, we also evaluate the similarity between reconstructed and ground-truth faces using SSIM. Results presented in Tab. 2 show that each progressive refinement leads to performance improvements, with our final version achieving the best results across all evaluation metrics.

4. CONCLUSIONS

We present a novel VFA approach for accurate and temporally stable video-based face alignment in this work. To begin, we use a coefficient-estimation model and a sampling approach to compute video-consistent identity and texture from a subset of frames. This allows us to create a realistic facial texture while also maintaining a consistent identity coefficient throughout the video. Following that, we use an iterative self-refinement process that employs differentiable rendering and optimization techniques to incrementally refine facial geometries and poses. Furthermore, the pre-constructed facial texture and identity coefficient are used to speed up face alignment in the remaining video frames. Experiment results validate the efficacy of our proposed method, demonstrating that it outperforms SOTA approaches in both accuracy and temporal stability.

5. ACKNOWLEDGEMENT

This work was supported by Key R&D Program of Zhejiang (No. 2023C01047) and the National Natural Science Foundation of China (Grant No. 62036010).

6. REFERENCES

[1] J Song and Z Liu, "Self-supervised facial action unit detection with region and relation learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2023, pp. 1–5.

[2] M Wei, W Zheng, Y Zong, X Jiang, C Lu, and J Liu, "A novel micro-expression recognition approach using attention-based magnification-adaptive networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 2420–2424.

[3] Y Feng, H Feng, M. J Black, and T Bolkart, "Learning an animatable detailed 3d face model from in-the-wild images," *ACM Transactions on Graphics*, vol. 40, no. 4, pp. 1–13, 2021.

[4] A Dapogny, K Bailly, and M Cord, "Decafa: Deep convolutional cascade for face alignment in the wild," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6893–6901.

[5] X Wang, L Bo, and L Fuxin, "Adaptive wing loss for robust face alignment via heatmap regression," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6971–6981.

[6] L Li, X Li, K Wu, K Lin, and S Wu, "Multi-granularity feature interaction and relation reasoning for 3d dense alignment and face reconstruction," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2021, pp. 4265–4269.

[7] J Guo, X Zhu, Y Yang, F Yang, Z Lei, and S. Z Li, "Towards fast, accurate and stable 3d dense face alignment," in *Proceedings of the European Conference on Computer Vision*, 2020.

[8] X Dong, S.-I Yu, X Weng, S.-E Wei, Y Yang, and Y Sheikh, "Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 360–368.

[9] P Micaelli, A Vahdat, H Yin, J Kautz, and P Molchanov, "Recurrence without recurrence: Stable video landmark detection with deep equilibrium models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22814–22825.

[10] X Dong, Y Yang, S.-E Wei, X Weng, Y Sheikh, and S.-I Yu, "Supervision by registration and triangulation for landmark detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3681–3694, 2020.

[11] C Zhu, H Liu, Z Yu, and X Sun, "Towards omni-supervised face alignment for large scale unlabeled videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 13090–13097.

[12] T Gerig, A Morel-Forster, C Blumer, B Egger, M Luthi, S Schönborn, and T Vetter, "Morphable face models-an open framework," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 2018, pp. 75–82.

[13] Y Han, Z Wang, and F Xu, "Learning a 3d morphable face reflectance model from low-cost data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8598–8608.

[14] Y Deng, J Yang, S Xu, D Chen, Y Jia, and X Tong, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[15] Z Wang, A. C Bovik, H. R Sheikh, and E. 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] D Kinga, J. B Adam, et al., "A method for stochastic optimization," in *International Conference on Learning Representations*, 2015, vol. 5, p. 6.

[17] A Bulat and G Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision*, 2017.

[18] J Shang, T Shen, S Li, L Zhou, M Zhen, T Fang, and L Quan, "Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency," *arXiv preprint arXiv:2007.12494*, 2020.

[19] H Jin, S Liao, and L Shao, "Pixel-in-pixel net: Towards efficient facial landmark detection in the wild," *International Journal of Computer Vision*, vol. 129, pp. 3174–3194, 2021.

[20] J Xia, W Qu, W Huang, J Zhang, X Wang, and M Xu, "Sparse local patch transformer for robust face alignment and landmarks inherent relation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4052–4061.

[21] J Shen, S Zafeiriou, G. G Chrysos, J Kossaifi, G Tzimiropoulos, and M Pantic, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 50–58.