# The 2nd 3D Face Alignment in the Wild Challenge (3DFAW-Video): Dense Reconstruction From Video

Rohith Krishnan Pillai
Robotics Institute
Carnegie Mellon University
rohithkp@cmu.edu

László A. Jeni
Robotics Institute
Carnegie Mellon University
laszlojeni@cmu.edu

Huiyuan Yang
Dept. of Computer Science
Binghamton University
hyang51@binghamton.edu

Zheng Zhang
Dept. of Computer Science
Binghamton University
zzhang@binghamton.edu

Lijun Yin
Dept. of Computer Science
Binghamton University
lijun@cs.binghamton.edu

Jeffrey F. Cohn
Department of Psychology
University of Pittsburgh
jeffcohn@pitt.edu

## Abstract

*3D face alignment approaches have strong advantages over 2D with respect to representational power and robustness to illumination and pose. Over the past few years a number of research groups have made rapid advances in dense 3D alignment from 2D video and obtained impressive results. How these various methods compare is relatively unknown. Previous benchmarks addressed sparse 3D alignment and single image 3D reconstruction. No commonly accepted evaluation protocol exists for dense 3D face reconstruction from video with which to compare them. The 2nd 3D Face Alignment in the Wild from Videos (3DFAW-Video) Challenge extends the previous 3DFAW 2016 competition to the estimation of dense 3D facial structure from video. It presented a new large corpora of profile-to-profile face videos recorded under different imaging conditions and annotated with corresponding high-resolution 3D ground truth meshes. In this paper we outline the evaluation protocol, the data used, and the results. 3DFAW-Video is to be held in conjunction with the 2019 International Conference on Computer Vision, in Seoul, Korea.*

## 1. Introduction

Performance of face alignment - the task of estimating detailed facial structure - has been improving steadily in the last decade, moving away from localizing a sparse set of 2D landmarks to estimating a dense 3D structure of the face.

Today, 3D face alignment technology serves many applications, from building personalized face rigs for animation to understanding nonverbal communication in real world conditions.

Previous datasets used for 3D face alignment research include multi-view 2D images annotated with sparse set of 2D landmarks (eg. Multi-PIE [9], CelebA[16], and AFLW[19]) and static 3D face scans accompanied with texture (eg. Bosphorus[21], BU3D[28], Stirling ESRC[1]).

3D face alignment from 2D video has been less studied than their single image based counterparts. Recent datasets, such as NoW[20], and MICC Florence [1], provide 2D video accompanied with 3D head scans for each subject including expressions. Many recent methods such as PRNet[5], 3DMM-CNN[25], MMFace[27] and Nonlinear 3DMM[24] use sparse 3D alignment and conduct single image 3D reconstruction. Such methods are popular due to availability of the large amounts of single image dataset, as opposed to video dataset built for such a task.

The majority of single image methods use a variant of 3D Morphable Model (3DMM) based fitting for 3D reconstruction including models such as Basel face model[8], Facewarehouse blendshapes[2], or FLAME[15].

Recently, differential renders have been used to train end-to-end deep learning based systems for 3D reconstruction, like in Mofa[23] and in GAN-FIT[7]. In contrast, multi-view or video based methods for 3D face reconstruction are less common. MVF-Net[26] introduced a method that takes multiple views of the subject to regress 3DMM parameters using an end to end CNN and minimizing a view-alignment loss. Similarly, [4] uses the complementary information from different images for shape aggregation to perform multi-image face reconstruction.

These multi-image based methods present good results for 3D face reconstruction, but how they compare to each other is relatively unknown. No commonly accepted evalu-

[1] http://pics.stir.ac.uk/

| Workshop/Challenge | Modality | Datasets | Ground Truth |
|---|---|---|---|
| 1st Workshop on 3D Face Alignment in the Wild (3DFAW) Challenge[11] | Single image, video | BP-4D-Spontaneous[32], BP-4DFE[31],Multi-PIE [9], Time-sliced[11] | 3D sparse mesh (Di3D, Strutcture-from-motion) |
| 1st 3D Face Tracking in-the-wild Competition[29] | Single image, video | 300W dataset[18], 3D Menpo database[30] | 3D sparse mesh (Deformable model fitting) |
| Workshop on Dense 3D Reconstruction from 2D Face Images in the Wild[6] | Single image | Stirling ESRC[1] , JNU 3D face dataset[14] | 3D dense mesh (Di3D, 3dMDface) |
| NoW challenge[20] | Single image | "not quite in-the-wild" (NoW dataset) [20] | 3D dense mesh (3dMDface) |
| 2nd 3D Face Alignment in the Wild Challenge (3DFAW-Video)[2] | Multiple images, video | 3DFAW-Video | 3D dense mesh (Di3D) |

Table 1. A summary and comparison of some of the challenges and workshops related to 3D face reconstruction in recent years.

ation protocol exists for dense 3D face reconstruction from video with which to compare them.

To enable comparisons, this paper introduces the 2nd 3D Face Alignment in the Wild (3DFAW-Video) benchmark. We created an annotated corpus of profile-to-profile videos obtained under a range of conditions: (i) high-definition in-the-lab video, (ii) unconstrained in-the-wild video from an iPhone device, and (iii) high-resolution 3D face scans from a Di3D imaging system. We also introduce a novel symmetric evaluation metric than the traditional 3D-RMSE scores to compare the various 3D reconstructions on the 3DFAW-Video dataset.

### 1.1. Previous related benchmarks

There have been a few prior workshops and competitions related to 3D face reconstruction in the recent past, Table 1 summarizes the most relevant ones. Some of the early challenges related to 3D face reconstruction consisted of the task of 3D face alignment from 2D video or even single image examples. The 1st Workshop on 3D Face Alignment in the Wild (3DFAW) Challenge[11] used the BP-4D-Spontaneous[32], BP-4DFE[31],Multi-PIE [9], and Time-sliced[11] datasets to provide the images and their respective 3D landmarks as ground truths, provided by an automatic algorithm. Similarly, 1st 3D Face Tracking in-the-wild Competition[29] also provided data from 300W dataset[18] and 3D Menpo database[30], focusing on the tracking of the 3D facial landmarks. They also used an automatic fitting algorithm to provide the ground truths for the data, like [11]. Since these competitions did not use real 3D scans for their ground truth, as [6] pointed out, these are not ideal for use in bench-marking, due to the limitations of the algorithms providing the ground truth making any learned technique at most as good as the ground truth algorithm.

Recent challenges such as the Workshop on Dense 3D Reconstruction from 2D Face Images in the Wild[6] mitigate this by using 3D dense mesh ground truths in the challenges to evaluate the various methods. This allows for very direct comparison of the accuracy of methods against the best estimate of the face shape of individuals. However, [6] did not provide any annotated 3D mesh ground truth dataset for training and allowed the use of any outside datasets, while using Stirling ESRC[1] and JNU 3D face dataset[14] containing 3D mesh ground truths only for the testing and validation phases respectively. The NoW challenge[20] introduced a new standard evaluation metric to the accuracy of the 3D reconstruction from single monocular images. Similar to previous challenges, the NoW challenge provided data in the form of images, and also improved on them by providing the 3D ground truth scans as a part of the "not quite in-the-wild" (NoW dataset) [20].

The 2nd 3D Face Alignment in the Wild (3DFAW-Video) Challenge[2] improves on the earlier 3DFAW challenge[11] by providing 3D dense mesh scans as the ground truth for quantifiable evaluation on the accuracy of differing methods in 3D reconstruction. Furthermore, unlike [6] and [20] we provide the 3D ground truth scans as well as their facial landmarks for the frontal images in the video for not only the testing and/or validation phases of the challenge but also for the training fold of the 3DFAW-Video dataset. Moreover, while most related datasets and challenges limit itself to just the single image modality, 3DFAW-Video provides video data for the reconstruction of the face. This allows for both multi-view and existing single image methods to be evaluated together on the same dataset and to be compared against the 3D dense mesh ground truth. Although, unlike the use of single images which can be in large numbers, datasets such as 3DFAW-Video that provide paired 3D facial scans and videos, tend to have smaller dataset sizes as a single video is recorded per subject compared to multiple images in an image based dataset. Nonetheless, the video data provided is richer in information given that the data inherently exposes multiple angles of the face allowing more

[2]http://3dfaw.github.io

information to be extracted regarding the 3D shape of the subject versus a single view image.

## 2. 3DFAW-Video Dataset

The 3DFAW-Video Dataset contains 3 different components for each of the 66 different subjects:

1. A high resolution 2D video from a DI3D system
2. Unconstrained 2D video from an iPhone
3. Hi-resolution 3D ground truth mesh

For both the high resolution 2D videos from DI3D 3D imaging system and the lower resolution 2D video from the iPhone 6 camera, the video captures an arc around the subject from profile-to-profile. The third component, the 3D reconstructed ground truth meshes for each of the subjects were created by merging multiple meshes from the DI3D 3D imaging system. Details of the data collection and 3D reconstruction of the ground truth meshes are described in greater details in the following subsections.

### 2.1. Data Acquisition

The DI3D 3D imaging system video was captured along with the corresponding 3D meshes in a controlled environment with 2 symmetric lights against a static dark background. The system consists of a RGB 2D color camera vertically flanked by a 3D sensor composed of a pair of stereo monochrome cameras. Dense passive stereo photogrammetry method is used to recover the 3D model for each frame, containing about 30k-50k vertices, and at a precision of 0.2mm RMS. The frame rate was set to 25fps, and each of the 2D texture images has a resolution of 1040x1392 pixels. The DI3D imaging system was fixed in place while the subjects were asked to sit in front of cameras and then slowly rotate their heads from left profile to right profile to capture a $5 \sim 10$s 3D sequence and the respective video.

The iPhone 6 video on the other hand was captured by a hand-held iPhone 6 in a much more 'in-the-wild' environment, although still indoors with ambient indoor lighting. Therefore, the iPhone 6 data provides more varied data, as it includes the subjects not being totally centered in various frames, and contains shaking that is characteristic of hand-held mobile video capture. Figure 1 shows the example of a video data from both the high resolution and the iPhone cameras provided.

The following subsections describe the procedure followed for the reconstruction of the ground truth meshes that provide the dense 3D reconstructions for each of the 66 subjects. The total n=66 subjects, were of the ages 18 to 28 with an average age of 19.74 and standard deviation of 2.3. The subjects are also racially and ethnically diverse as shown in Table 2, and also roughly balanced with respect to gender, with 36 females and 30 males. The subjects gave informed consent for the distribution and use of their video images for non-commercial research.

Figure 1. The montage shows selected frames of the profile to profile videos of one of the subjects, with the iPhone video on the top row, the high resolution video from the DI3D system on the middle row, and a few angles of the ground truth mesh on the bottom row.

| Ethnicity | No. of Subjects |
|---|---|
| African American | 1 |
| Asian | 20 |
| Latino/Hispanic | 7 |
| White | 35 |
| Others | 3 |

Table 2. The ethnic/racial distribution of the subjects in the 3DFAW-Video dataset.

### 2.2. Ground Truth Reconstruction

The data collected from using the DI3D system is in the form of a sequence of individual 3D meshes of roughly 20,000 vertices and 35-40k faces, covering a part of the face of the person. Each of the meshes are specified in the wave-front object format, with a corresponding image that is used for mapping the texture on to the mesh. Each of the mesh sequences contains approximately 130 meshes capturing the face from a front view of the face then panning to both the left and right profile views of the face and then back to the front view.

In order to reconstruct the ground truth 3D face scan from the multiple meshes that are included in a sequence, each of the sequence meshes are first cleaned manually us-

ing MeshLab [3], by deleting vertices and faces from the mesh at areas that have inaccurate projections. This is typically seen around the corners of a mesh where the occlusion or rapid change in depth lead to elongated projections of the mesh vertices. The ears, nose and neck are especially prone to casting wrongly projected vertices at their corners when the feature ends or occludes the points behind it. Care is taken to clean in order to minimize the number of wrongly projected vertices in the cleaned meshes in a sequence.

Once the cleaning of the sequences in the mesh is completed, then approximately 10 meshes are selected at key frames from the sequence including the mesh providing the frontal, left profile and right profile views and a few other views between them. Any mesh that contains features in the texture that could degrade the quality of the reconstructed mesh, such as frames where the subject blinks an eye, is dropped.

The selected 10 meshes from the sequence are then manually aligned and registered using CloudCompare[3]. The ten meshes are loaded and pairwise aligned starting with the frontal mesh and moving to consecutive meshes until reaching one of the side profiles. The same procedure is repeated for the meshes on the other side profile as well. The aligning of 2 meshes is conducted by point correspondence between the source mesh (the mesh that was previously aligned with the frontal mesh being the base case) and the target mesh. Approximately 5-6 correspondence points are used for the aligning procedure. In case the quality of the alignment was visibly inaccurate, more points were used for correspondence. However, this procedure only provides a rough alignment between the 2 meshes, and in order to align the two meshes more finely we use the iterative closest point (ICP) algorithm. The ICP is allowed to run until the RMSE difference is below $1\text{x}10^{-4}$ with a final overlap of 60-80%. The percentage of overlap is decided by the similarity between the 2 meshes. Again, the source and target meshes are the same as they were used in the rough alignment procedure. Once all the meshes have been aligned and registered using ICP in a pairwise manner, they are then merged together to create a new mesh.

The merged mesh from the registering of the meshes might have introduced some artifacts, especially around the borders of 2 aligned meshes or at areas where meshes weave and overlap each other. In order to reduce these artifacts and create a smooth watertight surface of the 3D scan we apply a Screened Poisson reconstruction on the merged mesh using the method described in [12]. This watertight surface mesh is used as the final reconstructed 3D ground truth for the subject as shown in Figure 2c.

[3]http://www.cloudcompare.org/

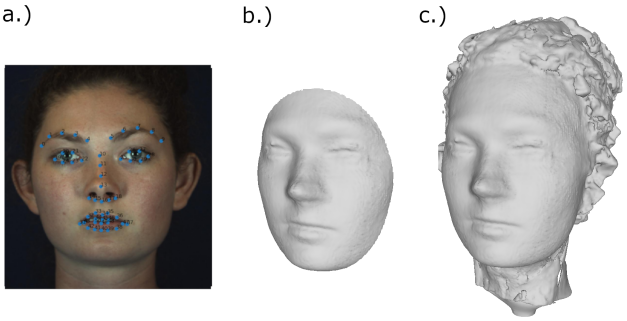


Figure 2. (a) The last 51 facial landmarks from dlib library. (b) The cropped 95mm cropped ground truth mesh. (c) The full watertight ground truth mesh for a sample subject.

| Data Fold | Total Meshes | Stratification | No. Subjects |
|---|---|---|---|
| Train | 26 | Both | 26 |
| Validation | 14 | HiRes | 7 |
| | | iPhone | 7 |
| Test | 26 | HiRes | 13 |
| | | iPhone | 13 |

Table 3. The number of subjects distributed across the multiple folds. The data includes the cropped ground truths and their 51 landmarks along with the videos in both high resolution and a lower iPhone captured video.

## 2.3. Data Folds

The 3DFAW-Video dataset is split into 3 subject-independent folds: the train, validation and the test set, as shown in Table 3. The training data fold contains both the high resolution (HiRes) video frames from the DI3D 3D imaging system as well as the iPhone 6 video for 26 subjects. We also provide 51 facial landmark annotations for each of the subjects with the frontal frames of each of the video sources, along with their corresponding 3D landmark locations. The 51 facial landmarks, as shown in Figure 2a, are the inner facial landmark subset of the 68 landmarks provided by the dlib library[13]. Since the challenge focuses on dense 3D reconstruction from 2D videos, the validation set provides only one video component, either iPhone 6 or high resolution (HiRes) frames. The validation set contains 14 total subjects with 7 videos of each resolution. The test set is similarly stratified to contain only one of the 2 resolution videos for each subject and forces the dense 3D reconstruction to be created from the specified sources. The third component, the 3D ground truth is provided for only the train fold and a smaller trimmed surface, as shown in Figure 2b, of the full watertight mesh is provided instead of the full watertight mesh. A distance of 95mm around the landmark at the tip of the subject's nose is used as the criterion for the trimmed mesh following the protocol in [4]. It is worth noting that the meshes

are non-textured and only provide the shape of the face and the corresponding 3DFAW-Video challenge is primarily interested in accurately capturing the face shape and hence did not need textured submissions. The data releases can be downloaded from the challenge website[4].

## 3. Evaluation Protocol

In order to evaluate the various different methods on the 3DFAW-Video dataset, we provide the evaluation code in Python for the competition. The evaluation protocol and code was adapted from the protocol described in [4]. The procedure includes the trimming of the predicted meshes for each subject to a radius of 95mm around their nose tip using the landmarks provided. The landmarks that are associated with each of the 51 facial fiduciary features are used to rough align the predicted meshes to their ground truths. Once the predicted meshes are trimmed to the region of the face corresponding to the ground truth meshes, they are then rigidly aligned using the iterative closest point (ICP) method. The aligned meshes are then compared using a metric that we introduce, using a sampled down copy of the mesh for efficiency. Both iPhone and HiRes videos based reconstructions are evaluated in the testing phase of the challenge due to the stratification of the video data in the test fold.

Although most 3D ground truth datasets commonly use 3D-RMSE as the metric for their evaluation, 3D-RMSE is not entirely a symmetric metric. Hence, for the 3DFAW-Video dataset we introduce, Average root mean square error(ARMSE), a modification of the 3D-RMSE metric that ensures that the metric is symmetric.

Average root mean square error, ARMSE, is an evaluation metric, which is the average of the point-to-mesh distance between the ground truth and predicted and vice verse. The Euclidean error is used as the distance metric between point and mesh, and is computed using the equation 1 below:

$$E(A_i, B) = min(\|A_i - B_v\|_2, \|A_i - B_e\|_2, \|A_i - B_f\|_2) \quad (1)$$

In equation 1, $A$ is the source mesh and $B$ is the target mesh. $A_i$ is a vertex in mesh $A$, $B_v$ is the closest vertex to $A_i$ on $B$, and similarly $B_e$ is the closest edge on $B$ to $A_i$, and $B_f$ is the closest face on $B$ to $A_i$. The above equation 1 calculates the shortest distance between a vertex in $A$ to the surface of the mesh $B$. The closest vertices, edges and faces on the target mesh $B$ are found using a nearest neighbor search. Here, $D(A, B)$ is defined between the 2 meshes, the source mesh $A$, and the target mesh $B$, with $N_a$ being the number of vertices in the source mesh $A$. The intermediate RMSE error, $D(A, B)$, from the source mesh $A$ to the target mesh $B$ can be calculated using the equation 2.

$$D(A, B) = \sqrt{\frac{1}{N_a} \sum_{i}^{N_a} E(A_i, B)^2} \quad (2)$$

$$ARMSE(X, Y) = \frac{100(D(X, Y) + D(Y, X))}{2I} \quad (3)$$

Then using this error metric $D(A, B)$, the RMSE scores between the predicted and ground truth and vice-versa are calculated. The 2 different RMSE scores are calculated with each of $X$,$Y$ meshes made to be the source mesh $A$ with the other as the target mesh $B$, because the nearest neighbor search for the closest vertices/edges/faces in the target mesh is not symmetric. This provides the 2 RMSE scores, $D(X, Y)$ and $D(Y, X)$ as in equation 3. Here, $I$ is the outer inter-ocular distance on the ground truth mesh $Y$, i.e. the euclidean distance between the $19^{th}$ and $28^{th}$ landmark points of the 51 dlib facial landmarks, in 3D space. The overall ARMSE score calculated by equation 3 is then scaled to report the error as a percentage of the outer inter-ocular distance of the subject.

The ARMSE metric allows for reducing the effect of the density of the predicted meshes from playing a very drastic role in determining the final score of the reconstruction unlike the traditional 3D-RMSE score. 3D-RMSE would produce low scores for a mesh that has fewer vertices, but lie close to the 'average face', as it only takes into account the one sided distance from the vertices of the predicted mesh to the surface of the ground truth mesh. By taking the reverse distance from the vertices of the ground truth to the surface of the predicted mesh, such 'average face' meshes would produce higher scores, and thereby penalize the mesh reconstruction score from the bias on mesh density. ARMSE, takes the average of both these scores and hence takes is less effected by the bias on density of the meshes than 3D-RMSE. The normalization by the outer inter-ocular distance is done to ensure that the inter-subject face size difference is captured in the metric. The ARMSE metric can be thought of as the average symmetric distance between the 2 meshes as a percentage of the subject's outer inter-ocular distance. Hence, the symmetrical ARMSE metric is an improvement over the traditional 3D-RMSE metric, by reducing the effects of density of meshes and also improving the readability of the metric. Although other metrics such as chamfer distance and earth mover's distance exists, the former of which is a bi-directional distance metric similar to ARMSE, they are primarily used for point clouds and not for 3D meshes and cannot capture the distance between points in one mesh to the mesh surface of another.

[4] https://3dfaw.github.io/

| Rank | Team | mean ARMSE |
|---|---|---|
| 1. | Zheng | 1.6962 |
| 2. | Shao et al.[22] | 1.8642 |
| 3. | Maldonado et al. [17] | 2.1429 |
| 4. | Chen | 2.1865 |

Table 4. The mean ARMSE scores of the different methods on the 3DFAW-Video test set.

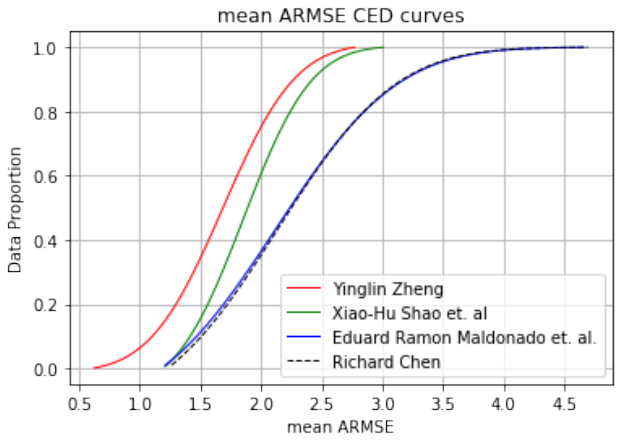


Figure 3. The cumulative error distribution of the participants' methods in the leader-board.

## 4. Summary of Competition Approaches

The competition was hosted on the Codalab[5] platform where eight teams submitted results. Of these eight teams, only four completed the challenge by submitting a technical description of their methods. Hence, we only report the results and methods of these four teams. Table 4 provides the final leader-board results for the competition, ranked by their mean ARMSE scores on the testing set of the 3DFAW-Video dataset. The methods used by the participants in the competition are discussed briefly below, from the ones that performed the best on the 3DFAW-video dataset onward.

Yinglin Zheng proposed a method that is based on the traditional 3D dense reconstruction using structure from motion(SFM), that is improved by utilizing facial prior knowledge. First, a face parser was used to segment the head region and SFM routine estimates the camera parameters and sparse 3D point cloud. Then the sparse point cloud is used for fitting a 3DMM, in the face subspace. Then a non-rigid registration is conducted to match the subspace shape more closely with the point cloud. A general dense reconstruction along with an outlier pruning by projecting the 3D points onto the union region of the parsing mask and projection region of the aligned shape is executed. Finally, a 2nd-pass non-rigid registration is conducted to align a template mesh with the pruned dense 3D point cloud by minimizing an energy function, to further enrich more details of the final mesh.

Xiao-Hu Shao et al.[22] used an ensemble of independent regression networks to create a framework consisting of multiple reconstruction branches and a subsequent mesh retrieval module. The multi-reconstruction branches produce 3D shapes using regression networks such as 3DDFA[33], PRNet[5], and MVFnet [26], and based on weighted linear combination all frames of the video. A synthesized image is rendered from the candidate shape, and texture maps predicted by each of the branches. Finally a mesh retrieval module selects the best fitting mesh using a weighted photo distance defined between the ground truth texture and the synthesized texture for the 3D shape. The training of their network was done on the 300W-LP dataset[33] which provides fitted 3DMM parameters for over 60,000 images.

Eduard R. Maldonado et al.[17] proposed a learning based method employing siamese neural networks that reconstructs 3D faces from either single or multiple images. Their multi-view solution uses the siamese networks to predict both the individual camera poses and the shape parameters of the 3DMM model for each of the multiple views. These shape parameters are then merged and then regressed on using another multi-layer perceptron to get the final shape parameters of the 3DMM model. The network was trained on their own dataset of 6,528 diverse individuals, and used a hyperparameter-less, unsupervised loss which is composed of the sum of the reprojection errors across all the different input views.

Richard Chen tried a much more traditional 3DMM fitting approach by finding a least squares solution for the PnP problem of keypoints, but augmented it by using race specific basis and mean face shapes. Most significantly, they used their own dataset to obtain the basis and mean face for eastern faces. They also used the basis and mean faces from the EOS project[10] and the Basel Face Model (BFM) [8]. Since the meshes from each model had different topologies, they were unified to single topology using a mesh transfer function. Finally, an exhaustive combination of the different basis and mean faces was used to then pick the reconstruction with the minimal error.

These were the methods used by some of the participants in the 3DFAW-Video challenge and it is apparent that 3DMM based approaches were unanimously the most common. The cumulative error distribution for the 4 different methods described above can be found in Figure 3.

## 5. Conclusion

In this paper we have presented the $2^{nd}$ 3D Face Alignment in the Wild from Videos (3DFAW-Video) Challenge dedicated to dense 3D face reconstruction from 2D video.

[5]https://codalab.lri.fr

The challenge evaluated the performance of the reconstruction on a new large corpora of profile-to-profile face videos annotated with corresponding high-resolution 3D ground truth meshes. The dataset included profile-to-profile videos obtained under a range of conditions: (i) high-definition in-the-lab video, (ii) unconstrained video from an iPhone device, and (iii) high-resolution 3D face scans from a Di3D imaging system. The challenge addresses the significant problem of reconstructing the dense 3D structure of the face from the two different video sources.

The paper reports results for four challenge participants that provided necessary technical descriptions of their methods. These results demonstrate room for potential improvements to be brought by future participants.

## 6. Acknowledgements

This work was supported in part by US Department of Defense grant W911SR-17-C0060, and NSF grants CNS-1629716 and CNS-1629898. We also thank Yu Deng and Jiaolong Yang from Microsoft Research Asia for providing the evaluation code and protocol for use in the challenge.

## References

[1] A. D. Bagdanov, A. Del Bimbo, and I. Masi. The florence 2d/3d hybrid face dataset. In *Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding*, J-HGBU 11, page 7980, New York, NY, USA, 2011. ACM.

[2] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013.

[3] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, and G. Ranzuglia. Meshlab: an open-source mesh processing tool. In *Eurographics Italian chapter conference*, volume 2008, pages 129–136, 2008.

[4] 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 Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[5] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 534–551, 2018.

[6] Z.-H. Feng, P. Huber, J. Kittler, P. Hancock, X.-J. Wu, Q. Zhao, P. Koppen, and M. Rätsch. Evaluation of dense 3d reconstruction from 2d face images in the wild. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 780–786. IEEE, 2018.

[7] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1155–1164, 2019.

[8] 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 (FG 2018)*, pages 75–82. IEEE, 2018.

[9] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[10] P. Huber, G. Hu, R. Tena, P. Mortazavian, P. Koppen, W. J. Christmas, M. Ratsch, and J. Kittler. A multiresolution 3d morphable face model and fitting framework. In *Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016.

[11] L. A. Jeni, S. Tulyakov, L. Yin, N. Sebe, and J. F. Cohn. The first 3d face alignment in the wild (3dfaw) challenge. In *European Conference on Computer Vision*, pages 511–520. Springer, 2016.

[12] M. Kazhdan and H. Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):29, 2013.

[13] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[14] P. Koppen, Z.-H. Feng, J. Kittler, M. Awais, W. Christmas, X.-J. Wu, and H.-F. Yin. Gaussian mixture 3d morphable face model. *Pattern Recognition*, 74:617–628, 2018.

[15] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017.

[16] Z. Liu, P. Luo, X. Wang, and X. Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15:2018, 2018.

[17] E. R. Maldonado, J. Escur, and X. Giro-i Nieto. Multi-view 3d face reconstruction in the wild using siamese networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.

[18] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18, 2016.

[19] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013.

[20] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7763–7772, 2019.

[21] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. Bosphorus database for 3d face analysis. In *European Workshop on Biometrics and Identity Management*, pages 47–56. Springer, 2008.

[22] X.-H. Shao, J. Lyu, J. Xing, L. Zhang, X. Li, X.-D. Zhou, and Y. Shi. 3d face shape regression from 2d videos with multi-reconstruction and mesh retrieval. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.

[23] A. Tewari, M. Zollhofer, H. Kim, P. Garrido, F. Bernard, P. Perez, and C. Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1274–1283, 2017.

[24] L. Tran and X. Liu. Nonlinear 3d face morphable model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7346–7355, 2018.

[25] A. Tuan Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5163–5172, 2017.

[26] F. Wu, L. Bao, Y. Chen, Y. Ling, Y. Song, S. Li, K. N. Ngan, and W. Liu. Mvf-net: Multi-view 3d face morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 959–968, 2019.

[27] H. Yi, C. Li, Q. Cao, X. Shen, S. Li, G. Wang, and Y.-W. Tai. Mmface: A multi-metric regression network for unconstrained face reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7663–7672, 2019.

[28] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 211–216. IEEE, 2006.

[29] S. Zafeiriou, G. G. Chrysos, A. Roussos, E. Ververas, J. Deng, and G. Trigeorgis. The 3d menpo facial landmark tracking challenge. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2503–2511, 2017.

[30] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen. The menpo facial landmark localisation challenge: A step towards the solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 170–179, 2017.

[31] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013.

[32] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[33] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.