Continuous Landmark Detection with 3D Queries

Prashanth Chandran Gaspard Zoss Paulo Gotardo* Derek Bradley
DisneyResearch|Studios
prashanth.chandran,gaspard.zoss,derek.bradley@disneyresearch.com, gotardop@gmail.com

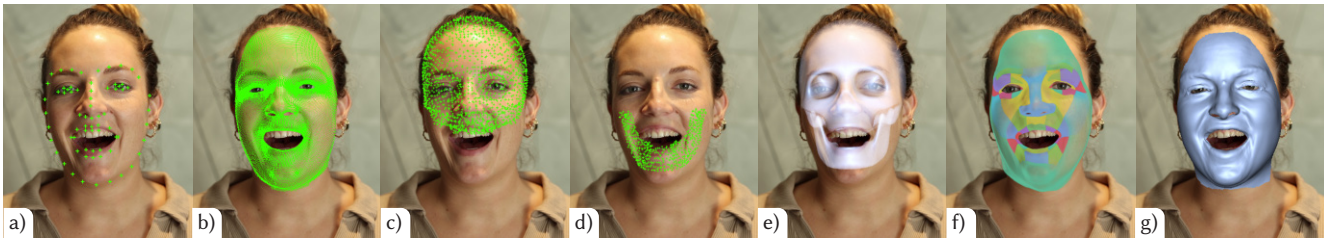

Figure 1. This paper presents a continuous landmark detector that is controlled intuitively with 3D queries. Once trained, a single model can be evaluated with the standard 68 landmarks layout (a), any arbitrarily dense layout (b), and can even be evaluated off-surface to get bone landmarks (c) and (d), which can be used to fit plausible skull and jaw meshes (e) to images. The same network can also be used to infer extremely dense landmarks, enabling applications such as face segmentation (f) or monocular performance capture (g).

Abstract

Neural networks for facial landmark detection are notoriously limited to a fixed set of landmarks in a dedicated layout, which must be specified at training time. Dedicated datasets must also be hand-annotated with the corresponding landmark configuration for training. We propose the first facial landmark detection network that can predict continuous, unlimited landmarks, allowing to specify the number and location of the desired landmarks at inference time. Our method combines a simple image feature extractor with a queried landmark predictor, and the user can specify any continuous query points relative to a 3D template face mesh as input. As it is not tied to a fixed set of landmarks, our method is able to leverage all pre-existing 2D landmark datasets for training, even if they have inconsistent landmark configurations. As a result, we present a very powerful facial landmark detector that can be trained once, and can be used readily for numerous applications like 3D face reconstruction, arbitrary face segmentation, and is even compatible with helmeted mounted cameras, and therefore could vastly simplify face tracking workflows for media and entertainment applications.

1. Introduction

Facial landmark detection has become extremely popular in computer vision and graphics applications. In particular, many applications in visual effects such as 3D facial reconstruction, tracking, face swapping and re-enactment rely on accurate facial landmark detection as one of the first steps in the process. It is therefore a crucial task and has been studied extensively for the past several decades, and the field has seen immense progress thanks to advances in deep learning.

State-of-the-art solutions for facial landmark detection are based on neural networks, and they operate by training the network to predict a fixed set of landmarks, leveraging large datasets of hand-annotated images. Most standard algorithms predict a set of 68 sparse landmarks spread across the face (Fig. 1 (a)), in a very specific and predefined layout [33]. However, recent work has shown that predicting denser landmarks on the face is better for tasks like face reconstruction [43]. This brings up the question of how many landmarks one has to predict for optimal performance (depending on the application), and what is the preferred layout of these landmarks? One of the biggest issues with traditional facial landmark detectors is that you need to decide on the number and layout of the landmarks ahead of time, then obtain annotated data with the corresponding landmarks and ultimately train the detector. Later, at runtime, the landmark layout cannot be changed.

An ideal landmark detector would not be bound to a specific fixed landmark layout. Such a detector could be trained once and then used in several applications with different landmark configurations. For example, in face image segmentation you may want to track landmarks corresponding to segment boundaries, but then in 3D reconstruction you might want to track landmarks corresponding to the vertices of your 3D face mesh. For individuals with specific face details like freckles or moles, you might want a de-

[0]Now at Google

tector that can track these user-defined points for the application of digital video touchup. For each application, with today's detectors you would need to train separate landmark detection networks, one for each landmark layout.

In this work we aim to reformulate how landmarks have conventionally been predicted with neural networks. We propose novel architectures for continuous, unlimited landmark detection at runtime. In other words, our method allows for an arbitrary number of landmarks to be predicted in any layout at inference time without retraining. As such, we propose the ideal landmark detector for multiple application use. The design of our method is simple, combining an image feature extractor with a queried landmark predictor; the latter takes the image descriptor together with a 3D query point relative to a template 3D face surface and predicts the corresponding 2D landmark location in the image. Since the 3D query points can be arbitrary, the result is continuous and unlimited landmark detection (Fig. 1 (b)). As our approach is modular, we evaluate multiple architecture options for both the feature extractor and the queried predictor, allowing different designs that tradeoff accuracy and runtime. Furthermore, we will show that the query points do not even need to lie on the surface of the template face, allowing to predict 2D landmarks for volumetric features on the skull, jaw, teeth or eyes (Fig. 1 (c) and (d)).

In addition, an important benefit of our design is that we do not need to have training data with a single dedicated number of landmarks in a specified layout on the face. This means that our architecture can leverage multiple different pre-existing datasets at training time, even if they do not have consistent annotations. This fact, combined with the beauty of specifying any landmark layout at runtime makes our continuous landmark detector powerful, with applications in several areas of face capture including reconstruction, tracking, segmentation (Fig. 1 (e)-(g)) and many others. As a summary, we can enumerate the main benefits of our new landmark predictor as follows:

- Our method offers the ability to predict any desired landmark on the face at inference time without retraining the network.
- We can track non-standard landmarks like pores, moles or dots drawn on the face without training a specific predictor.
- Our method is not restricted to the face surface, and allows to predict landmarks for volumetric features like the skull, jaw, teeth and eyes.
- The size of the neural network is agnostic to the number of output landmarks.

2. Related Work

Nowadays, detecting facial landmarks is almost always solved with deep learning. While generally accurate on frontal views, traditional non-deep methods such as cascade regressions [7, 49] are usually outperformed by their modern counterparts. We refer the reader to the excellent surveys of Wang et al. [39] and Wu and Ji [48] for a thorough review of those traditional methods. In the following we will summarize the most relevant works but refer to the recent survey of Khabarlak and Koriashkina [22] for a more in depth summary of the field.

Deep learning based methods generally leverage a backbone to extract relevant image features which are then used to derive the landmarks. The choice of one over another is often dictated by memory usage and the desired speed or accuracy of the predictions [14, 16, 22, 54]. Often several choice of backbones are proposed [13, 18, 28, 50], each with their own advantages and disadvantages. One can split the methods predicting landmarks in two categories. So called *direct prediction* methods [13, 19, 28, 42, 43, 50, 55, 57] which regress the landmarks coordinates directly and *heatmaps prediction* methods [4, 5, 8, 23, 38], which first predict the distribution of each landmark, followed by an *argmax* operation to extract its location. The method we propose falls in the first category and directly predict the landmark coordinates on the image plane without using heatmaps.

By design of their network architecture, the vast majority of existing methods can only predict a specific set of landmarks. The number of channels of the predicted heatmaps tensor or the width of the last linear layer regressing the landmarks coordinates dictates how many landmarks are predicted. Moreover, the datasets used to train existing methods typically have a fixed set of sparsely annotated landmarks ranging from 21 to 98 [6, 24, 33, 42, 46]. This forces existing methods to only train on datasets with compatible landmarks layout. Some works have attempted to alleviate this drawback: *Dense Face Alignment* [30] follows the steps of more traditional methods and predicts landmarks using a deformable 3D face model. It uses the annotated training landmarks as constraints on the mesh, thus allowing training with arbitrary layouts. Similarly, *LDDMM-Face* [51] shows limited cross-annotation results by estimating shape model deformations. Finally, *Look at Boundary* [46] uses an intermediate face feature boundary heatmap together with a direct prediction module *per* layout, allowing to train the method on several datasets, with a dedicated head for each one. In contrast with these methods, we do not rely on specific layouts for training nor rely on having camera estimates, furthermore a user can arbitrarily change the queries at inference with our method.

As shown in Section 4, our method also allows for extremely dense queries, allowing to predict per pixel uv-coordinates or face segmentation. These dense predictions are typically acquired using specialized networks [12, 15, 32, 34–36], but our method allows a direct extension to face segmentation within our landmark estimation framework.

Over the last few years, Transformer architectures were used in the context of landmark prediction, either as the feature extraction backbone or as the direct prediction head [26, 27, 41]. One of the variants proposed in this paper also leverages Transformers but takes inspiration from works that encode the shape queries on a canonical volume [10, 11, 29] as we describe in Section 3.

In summary, despite a large body of work in facial landmark detection, we present the first method that can be trained once on any combination of datasets with different landmark configurations, and then allows to predict any sparse or dense landmark set at runtime, including off-surface features like on the skull, jaw, eyes and teeth.

3. Continuous Landmark Detection

We propose a novel, modular framework for continuous unlimited landmark detection on faces. Our neural solution primarily consists of two components; i) an image feature extraction network $\mathcal{F}$, and ii) a queried landmark predictor $\mathcal{P}$ (see Fig. 2). The feature extraction network $\mathcal{F}$ takes a 2D image of a face as the input and yields an image descriptor. The queried landmark predictor $\mathcal{P}$ combines the information from the image descriptor and a collection of 3D queries defined on a canonical face shape C to predict locations corresponding to the queried landmarks in screen space. We will now describe the details of both these components and offer insights into how this simple formulation substantially increases the flexibility and power of 2D landmark predictors. Note that we purposely keep our algorithm description at a high level in Section 3.1 and Section 3.2, and then offer several specific implementation options in Section 3.3, which we evaluate in Section 4.

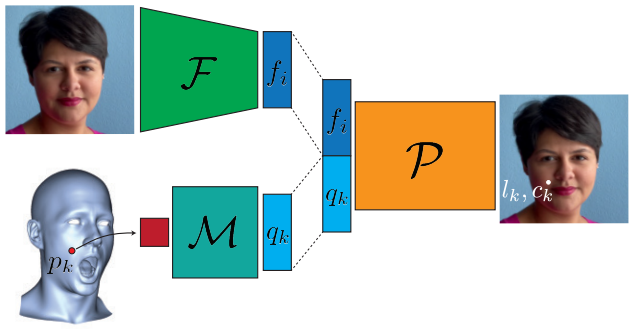

Figure 2. Here we show how a single 3D query point on the face (p_k) results in a corresponding 2D landmark on the image (l_k) with corresponding confidence (c_k).

3.1. Image Feature Extraction

The input to our feature extraction network $\mathcal{F}$ is an image $\mathcal{I}$ of a face. We apply a traditional normalization strategy [20] to ensure that the face is centered in the image and normalized in position, orientation and scale. This normalization step requires the detection of four face landmarks, two for the eyes and two for the mouth. Although the entire goal of our work is to predict face landmarks in images, we recognize that sparse landmark detection is a well studied problem and we leverage an off-the-self predictor [3] to detect these four landmarks used only for normalization.

The feature extraction network outputs a d-dimensional image descriptor f_i, specifically

$$f_i = \mathcal{F}(\mathcal{I}) \in \mathcal{R}^d, \tag{1}$$

where $d = 768$ in practice. The image descriptor f_i will be used by the queried landmark prediction network $\mathcal{P}$ to output the final 2D landmarks, as we describe next.

3.2. Queried Landmark Predictor

The second component of our architecture is a queried landmark predictor $\mathcal{P}$ which operates on the descriptor f_i and point samples from a canonical shape C. The canonical shape is a fixed template face from which we can sample 3D position queries that implicitly map to the 2D landmarks we wish to predict. Specifically, any 3D position p_k corresponding to a desired output landmark l_k is sampled from C and position encoded to obtain query $q_k \in R^B$. We fix $B = 64$ for all our experiments and for position encoding we use a 2-layer MLP $\mathcal{M}$ to learn the mapping from p_k to q_k during training. The function of the positional queries q_k is to inform $\mathcal{P}$ about the landmarks it needs to predict from the feature descriptor f_i for image $\mathcal{I}$. For example, if a user would like to predict the location of the nose tip in the input image $\mathcal{I}$, the query point q_j would be derived from the 3D point p_j corresponding to the nose tip on the canonical shape C. Note that the canonical shape already implicitly serves as a spatial prior over the distribution of facial landmarks in the output. We will show in Section 4.5 that our method allows to sample 3D points even off the surface of the canonical shape C, yielding 2D landmark tracking for volumetric objects like bones.

For decoding 2D landmarks, a positional query q_k is concatenated with the image descriptor f_i and fed as input to $\mathcal{P}$. Note that if multiple output landmarks have to be predicted for the same input image, the image descriptor f_i is duplicated n times (given n landmarks) and concatenated with the n queries $[q_0, q_1, ...q_{n-1}]$. In other words, the image descriptor f_i remains the same irrespective of the landmarks predicted on the output side. This disentanglement is a crucial property of our design that encourages the feature extraction network $\mathcal{F}$ to learn a very expressive feature space. Taking the concatenated data as its input, the queried landmark predictor P yields a 2D landmark l_k and a scalar confidence value c_k [43], which indicates how confident P is about the predicted landmark location. Mathematically speaking,

$$q_k = \mathcal{M}(p_k),$$
$$l_k, c_k = \mathcal{P}(f_i, q_k), \tag{2}$$

where $f_i \in \mathcal{R}^d$, $q_k \in \mathcal{R}^B$, $l_k \in \mathcal{R}^2$ and $c_k \in \mathcal{R}^1$.

As our network only predicts the 2D position of the 3D query from the canonical shape, our architecture has several advantages. The first is that it allows us to mix and match datasets with inconsistent annotations as long as the annotations from the individual datasets have a corresponding query point in a canonical space. Note that these queries have to be defined only once per dataset (irrespective of the size of the dataset) and are therefore very inexpensive to annotate. The second is that since the landmark predictor operates on a per-query basis, the size of the network does not grow with the number of output landmarks. If a user is interested in predicting multiple landmarks corresponding to multiple queries, these queries can simply be stacked along the batch dimension to predict multiple landmarks for the input image in parallel. We also note that one could also choose to operate in an alternate canonical space (UV domain for instance) and continue to retain all the benefits of our method.

3.3. Implementation Details

For the feature extraction network, we propose to use the recently developed ConvNext encoder [31], which is a convolutional encoder that has achieved almost state-of-the-art performance on several image processing benchmarks. However the modular nature of our architecture allows this feature extractor $\mathcal{F}$ to be replaced with any other network depending on the user's application scenario. For example, it might be interesting to consider lightweight feature extraction backbones such as MobileNetV3 [17] to achieve real time performance on mobile devices. We evaluate both ConvNext and MobileNetV3 in Section 4.7.

Similar to the modular nature of the feature extraction network $\mathcal{F}$, we also propose two different design choices for the queried landmark predictor $\mathcal{P}$. In particular, we evaluate i) a 4 layer MLP that operates on individual queries, and ii) a transformer [37] that can operate on a sequence of queries at once, leveraging self-attention to correlate landmark positions in the output. We show a quantitative evaluation based on both accuracy and performance of these architectural options in Section 4.7, and we provide a detailed explanation of the architecture variants in the supplemental document.

Both the feature extraction network $\mathcal{F}$ and the queried landmark predictor $\mathcal{P}$ are trained end-to-end using a gaussian log likelihood loss function in a supervised fashion, similar to Wood et. al [43], wherein the landmarks are normalized in screen space to stay in the range of [-1, 1]. Such a formulation allows for errors in the output as long as the network is not confident about such landmarks.

3.4. Training Data

We train our method on a multitude of datasets including sparse and dense facial landmarks, both in-the-wild and in studio, and also off-surface landmarks corresponding to anatomical structures like eyes, teeth, skull, and jaw.

For facial surface landmarks, we use 300-W [33] and the *Fake-it-till-you-make-it* [42] dataset, both containing sparse landmark annotations in-the-wild, and a studio dataset consisting of dense landmark annotations ($\approx$ 50000 landmarks per image). The dense landmark dataset is obtained from a multi-view 3D face database [9], where dense 3D surface points are projected into the multi-view face imagery. Note that we generally train with a mix of sparse and dense annotations as the dense annotations from the studio dataset provides information about how to spatially interpolate landmarks for in-the-wild imagery. That said, we also provide an evaluation of performance when training on different subsets of the data in Table 1 (Section 4). Furthermore, several publicly available alternatives to the studio data we use are possible, for example FaceScape [52], which we evaluate in the supplemental document.

As mentioned earlier, our formulation allows for querying 3D points anywhere in the volume around the canonical face C, allowing to train on unique datasets that contain tracked 2D positions for non-standard landmarks, such as on eyes, teeth or even on bones inside the head. In order to obtain tracked landmarks for these objects to use for training data, we turn to recent digital human reconstruction methods that have focused specifically on reconstructing subject-specific teeth geometry [44], estimating rigid skull motion [1,45], and tracking rigid jaw motion [58]. Applying these techniques on the multi-view 3D face database mentioned earlier [9] yields the ability to obtain anatomical landmarks for training our predictor. In these cases, the 3D geometry of the teeth, skull and jaw shapes in canonical space allow to naturally sample query points, and the corresponding projections of those points in the multi-view imagery provides the ground truth 2D landmarks for training. Furthermore, we extend our dataset with sparse eye landmarks by utilizing DatasetGAN [53], which allows to efficiently obtain annotated landmarks on synthetic face images with minimal human effort. Here we annotate 10 eye landmarks (5 on each eye) on 20 synthetic face images generated by StyleGAN2 [21], and apply DatasetGAN to automatically annotate an additional corpus of synthetic images. The query points for the eye landmarks are also manually annotated (one time) in 3D space relative to the canonical face mesh C.

Fig. 3 shows a very small overview of our entire dataset. In total we train with the 3129 face images containing 68-landmark annotations from 300-W, 100000 images with 70-landmark annotations from *Fake-it-till-you-make-it*, 100000 face images with dense ($\approx$ 50000) landmark annotations

from the studio dataset, a subset of 25000 studio images with annotated skull (1000 landmarks) and jaw (500 landmarks), a subset of 3000 with annotated teeth landmarks (500 for upper teeth, 500 for lower teeth), and finally 3000 synthetic images with 10 eye landmarks each.

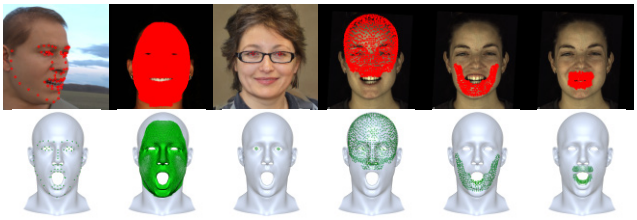

Figure 3. Our method can leverage a collection of different landmark datasets with inconsistent annotations. Here we show a small overview of the 2D landmarks (red, top row), along with the corresponding 3D queries (green, bottom row). From left to right: sparse in-the-wild landmarks [42], dense studio landmarks [9], eye landmarks [53], skull landmarks [1, 45], jaw landmarks [58], and teeth landmarks [44].

To regularize the training of the model, we perform color-space augmentations (eg. hue, saturation, brightness, contrast, pixel jitter) and geometric augmentations (integral and fractional shifts, isotropic and anisotropic scaling, rotation) of the image and the ground truth landmarks, to make our model robust to these effects at runtime. Although we train the network using discrete landmark locations on the face, the smooth nature of the predictor will allow any (in-between) landmark to also be queried accurately.

4. Results

We now present the results of our method and offer various evaluations and ablation studies. Please refer to the supplemental video and document for additional details and evaluations.

4.1. In-the-Wild Landmark Detection

We show the ability of our method to predict both sparse and dense 2D landmarks for in-the-wild test images in Fig. 4. Note that the same pre-trained network is used at inference time to predict all of the layouts shown in Fig. 4. For predicting different landmark layouts on the same image, only the 3D canonical queries need to be modified and the image feature extraction network needs to be evaluated only once. Due to the fact that we can train on many different datasets, the extensive variety of data available to our network allows our predictor to even work well in the complex situation of helmet-mounted camera imagery (Fig. 5), where dot tracking is often used for facial motion capture in digital productions. Our approach could ultimately remove the requirement for dot makeup.

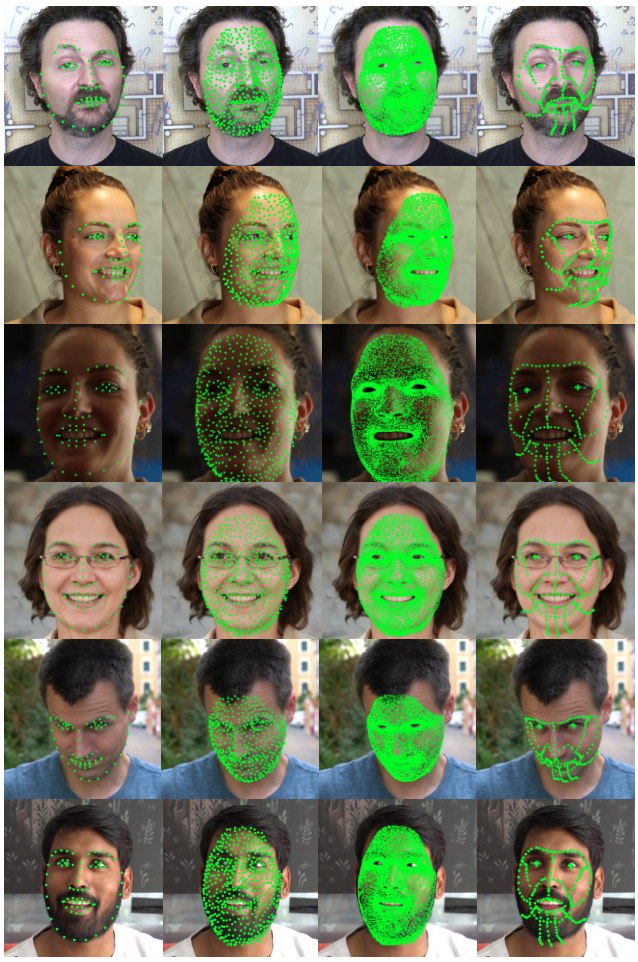

Figure 4. Our method can predict arbitrary landmark layouts on faces in-the-wild at runtime without re-training.

Figure 5. Our trained network can also be used for the complex situation of helmet-mounted camera images, such as often used in facial performance capture in film productions.

We next evaluate our landmark predictor quantitatively on the 300-W challenge dataset [33]. As described in

Section 3, the modular nature of our method naturally allows for multiple choices for the feature extraction network $\mathcal{F}$ and the query prediction network $\mathcal{P}$. In our work, we evaluate two choices for the feature extraction network $\mathcal{F}$: i) a ConvNext encoder [31], and ii) MobileNetV3 [17], and two choices for the query prediction network $\mathcal{P}$: i) a 4 layer MLP, and ii) a 4 layer Transformer MLP with 8 self attention heads (please see the supplemental document for exact network architecture details). As such, we could consider four different combinations of architectures (ConvNext+MLP, ConvNext+Transformer, MobileNet+MLP, MobileNet+Transformer). As the primary motivation for using the MobileNet backbone is speed and transformers are traditionally slow, we leave out the configuration combining a MobileNet backbone with a transformer query predictor from our evaluations. Additionally, the ability of our method to leverage multiple training datasets, even with inconsistent annotations for training adds another dimension to consider for quantitative evaluations. Here, we again consider two scenarios: i) training only on a synthetic dataset with 70 sparse annotations [42], which we refer to in the evaluation as the *Fake-it* dataset, ii) training on the full dataset described in Section 3.4 with a mix of in the wild sparse ground truth and dense annotations from studio captures for the skin and facial anatomy (referred to as *all data*). Table 1 presents the results of the evaluation on the *300 faces in the wild* challenge [33]. We report the normalized mean error on the 300-W Common and Challenging subsets even though our method was not trained on this dataset. We demonstrate comparable performance to state-of-the-art methods trained on the 300-W dataset, while outperforming Wood et al. [43], which is the method closest to ours in terms of predicting dense face landmarks. We further note that unlike Wood et al., we do not perform label translation to account for the differences in annotation between the training and evaluation datasets. In Table 4, we also show results of training our ConvNext+MLP variant only on the 300-W training set.

We also evaluate dense landmark prediction in the form of dense normalized mean error (NME) on left-out samples from the 3D dataset of Chandran et al. [9], for the different versions of our network in Table 2. Our queried landmark predictor allows predicting interpolated landmarks, even when trained on the sparse 70-landmark *Fake-it* dataset, as shown in Fig. 6 ((a) and (b)). Note, however, that in this case the method does not extrapolate well to dense landmarks given the sparse training data (Fig. 6 (c)). Only when trained on the full dataset including dense landmarks does our method produce accurate dense predictions in the wild (Fig. 6 (d)).

4.2. Query Optimization

In addition to specifying landmarks to predict as 3D query points, another use case is when a user defines one or more 2D points on a face image that they wish to track. For example, an artist may specify particular points on a person's face like moles or blemishes that they wish to paint out of a video. In these cases, our method allows to optimize for the corresponding 3D query point corresponding to each desired 2D image pixel. Using those query points for prediction, our network can track the specified face pixels over any video or additional image of that person. We show such an example of tracking a user-defined face point in Fig. 7, and compare the tracking to traditional optical flow [2]. Notice how our predictor can handle frames where the face point is occluded, which tends to cause tracking failures for optical flow.

A collection of queries can also be optimized at the same time. We demonstrate a practical application where such a feature might be necessary for consistent dot placement for

Table 1. Quantitative evaluation on the 300-W challenge.

Method (*Dataset*)	**Common Subset**	**Challenging Subset**
LAB [47] (*300-W*)	2.98	5.19
AWING [40] (*300-W*)	**2.72**	4.52
ODN [56] (*300-W*)	3.56	6.67
LUVLi [25] (*300-W*)	2.76	5.16
Wood et. al [43] (*Fake-it*)	3.03	4.80
Ours		
ConvNext+MLP (*Fake-it*)	3.17	5.96
MobileNet+MLP (*Fake-it*)	3.62	7.85
ConvNext+Transformer (*Fake-it*)	2.99	4.71
ConvNext+MLP (*all data*)	3.16	4.46
MobileNet+MLP (*all data*)	3.20	6.14
ConvNext+Transformer (*all data*)	2.87	**4.42**

Table 2. Quantitative evaluation on dense 2D landmarks.

Model (*Dataset*)	**Dense NME**
ConvNext+MLP (*all data*)	1.4
MobileNet+MLP (*all data*)	1.95
ConvNext+Transformer (*all data*)	1.07

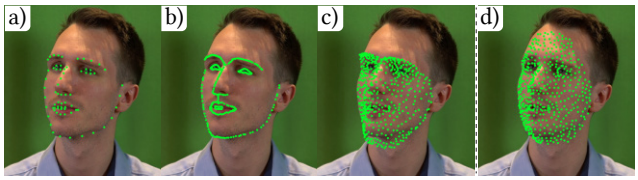

Figure 6. We show the performance of our method when trained solely on the *Fake-it* dataset [42] and querying a) the 70 landmarks seen at training time, b) interpolated positions between the training landmarks and c) dense 500 landmarks (failing). In d) we show a comparison of querying the same dense landmarks but trained on our full dataset as described in Section 3.4. While training with only 70 landmarks allows querying interpolated positions, only the model trained on the full dataset allows to query dense landmarks.

Figure 7. Our method allows to track user-defined features across a video, as shown by the pink dot on the eyebrow. In contract to optical flow tracking [2], our method is not affected when the face point is occluded.

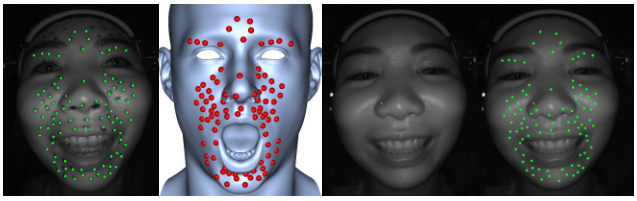

Figure 8. Query optimization for markers in an HMC image. Left to right: Single frame annotation of markers, optimized queries on the canonical shape, a new frame without dot makeup, and predicted markers on that frame to aid consistent dot placement.

helmet-mounted camera (HMC) face tracking in Fig. 8. By annotating a single frame containing makeup dot positions, our method can track facial markers through the remaining video. Furthermore, to aid consistent marker placement, we show the predicted landmarks on the same subject without markers captured during a different session.

4.3. Face Segmentation

Our method can also be used in facial segmentation tasks, where the goal is to divide a face image or video into different pre-defined regions (e.g. the nose, lips, eyes, etc). Usually facial segmentation is solved using specialized networks. However the ability of our method to predict arbitrarily dense landmarks extends it for the purpose of face segmentation as well. A user can further segment the face into multiple layouts at inference time, as shown in Fig. 9.

4.4. Artistic Texture Editing

Our method enables new possibilities in 2D face editing by allowing applications such as video face painting without requiring an explicit 3D reconstruction of the face. In Fig. 10 we show how a tattoo can be propagated across multiple identities, expressions and environments in a consistent manner. This can be achieved by simply annotating or designing the tattoo on the canonical shape and query these positions from our model.

Figure 9. Face segmentation in arbitrary layouts is also enabled by our method by predicting dense landmarks for each segment class on the query shape.

Figure 10. Our method can readily be used to artistically edit face images without the need for 3D face reconstruction.

4.5. Facial Anatomy Tracking

As described in Section 3.4, we can also leverage anatomy datasets for training. While not producing anatomically accurate results, our method can provide plausible, temporally smooth 2D landmarks which can be used to rigidly track 3D facial anatomy, as shown in Fig. 11. Furthermore, by adding eye landmarks to the training data, our method can just as well predict eye landmarks in the wild, as shown in Fig. 12. Please see the supplemental content for more examples including teeth landmark prediction.

4.6. Facial Performance Capture

Since our method can predict an arbitrarily dense number of landmarks, it can also be useful for face reconstruction in 3D. Fig. 13 shows the result of fitting an actor-specific face model (defined by Wu et al. [45]) to the landmarks predicted by our method. Using the monocular reconstruction method of Wu et al. [45] as a baseline, we present mean, median and standard deviation of 3D reconstruction errors for different landmark configurations in Table 3. The results in Fig. 13 correspond to the last row of Table 3, which produce the lowest errors. Please refer to our supplemental video for more 3D reconstruction results.

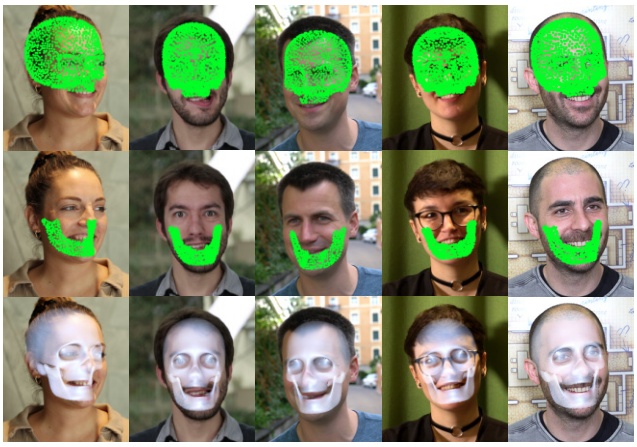

Figure 11. We show how our method can predict non-standard volumetric landmarks corresponding to skull and jaw features (top and middle row), and these can be used to fit anatomical geometry (bottom row).

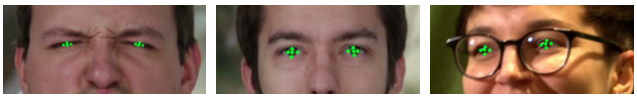

Figure 12. Our method can predict high quality eye landmarks, which could be used in eye tracking applications.

Table 3. Quantitative evaluation on 3D face reconstruction.

Ours	**Mean error (mm)**	**Median error (mm)**	**Std**
68 landmarks	2.88	2.48	1.94
500 landmarks	2.63	2.07	1.95
50K landmarks	2.34	1.74	1.93
68 landmarks, 3 views	2.76	2.14	2.06
500 landmarks, 3 views	2.26	1.73	1.86
50K landmarks, 3 views	2.13	1.59	1.76

Figure 13. Using dense landmark prediction, our method can be used for high quality 3D facial performance reconstruction.

4.7. Ablation Studies

Runtime Analysis We analyze the inference time of our three architectural variants in Fig. 14. Since the number of decoded queries also influences run-time, we plot the speed of inference in frames per second for different architectural choices. As the number of queries increases, we see that the transformer variant drops in run-time performance significantly, while the MLP variants retain real time or interactive frame rates. The numbers were computed on a Nvidia 1080Ti GPU and were obtained by averaging 1000 runs.

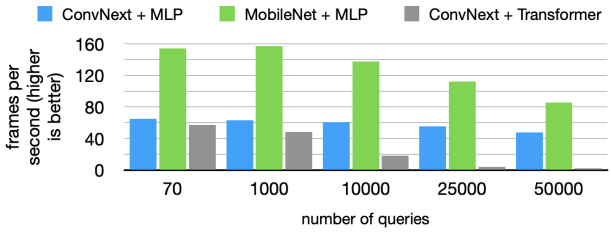

Figure 14. Run time vs. Number of decoded queries for each of our network variants. Our ConvNext + MLP variant offers a middle ground between performance and accuracy.

Table 4. Effect of color and geometric augmentations while training on the 300-W [33] train set and evaluating on the 300-W *Common* test set.

Method	**NME**
ConvNext+MLP (*No Augmentations*)	4.27
ConvNext+MLP (*Color only*))	4.24
ConvNext+MLP (*Affine only*)	3.87
ConvNext+MLP (*Color + Affine only*)	3.63

Effect of augmentation Similar to many state-of-the-art landmark prediction networks, we employ image space augmentations (in color and geometry) on the training data to make our models robust to common image degradations. In Table 4, we demonstrate the effect of these augmentations on one of our architectural variants on the 300-W dataset.

5. Conclusion

In this work we present a new formulation for facial landmark detection based on 3D query points. For the first time, our method allows to predict arbitrary landmark locations at runtime without re-training the network. Additionally, our approach allows to leverage a multitude of existing training datasets, even with inconsistent annotations. We propose a simple network design that is modular and allows different architecture variants with tradeoffs of speed and accuracy. Our method can predict sparse or dense landmarks, both on the face and off surface, allowing applications beyond traditional landmark detection like face segmentation, artistic texture editing, facial anatomy tracking, performance capture, and user-specific landmark tracking.

References

[1] Thabo Beeler and Derek Bradley. Rigid stabilization of facial expressions. *ACM TOG*, 33(4), 2014. 4, 5

[2] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, pages 25–36, 2004. 6, 7

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230, 000 3d facial landmarks). In *ICCV*, pages 1021–1030, 2017. 3

[4] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017. 2

[5] Adrian Bulat and Georgios Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *CVPR*, June 2018. 2

[6] Xavier P. Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *ICCV*, pages 1513–1520, 2013. 2

[7] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. In *CVPR*, pages 2887–2894, 2012. 2

[8] Prashanth Chandran, Derek Bradley, Markus Gross, and Thabo Beeler. Attention-driven cropping for very high resolution facial landmark detection. In *CVPR*, 2020. 2

[9] Prashanth Chandran, Derek Bradley, Markus Gross, and Thabo Beeler. Semantic deep face models. In *3DV*, pages 345–354, 2020. 4, 5, 6

[10] Prashanth Chandran, Gaspard Zoss, Markus Gross, Paulo Gotardo, and Derek Bradley. Facial Animation with Disentangled Identity and Motion using Transformers. *Computer Graphics Forum*, 2022. 3

[11] Prashanth Chandran, Gaspard Zoss, Markus Gross, Paulo Gotardo, and Derek Bradley. Shape transformers: Topology-independent 3d shape models using transformers. 41(2):195–207, 2022. 3

[12] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, September 2018. 2

[13] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *CVPR*, June 2018. 2

[14] Pengcheng Gao, Ke Lu, and Jian Xue. Efficientfan: Deep knowledge transfer for face alignment. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, ICMR '20, page 215–223, 2020. 2

[15] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 2

[16] Xiaojie Guo, Siyuan Li, Jiawan Zhang, Jiayi Ma, Lin Ma, Wei Liu, and Haibin Ling. PFLD: A practical facial landmark detector. *CoRR*, abs/1902.10859, 2019. 2

[17] Andrew Howard, Ruoming Pang, Hartwig Adam, Quoc Le, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, and Yukun Zhu. Searching for mobilenetv3. In *ICCV*, pages 1314–1324, 2019. 4, 6

[18] Haibo Jin, Shengcai Liao, and Ling Shao. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *Int. J. Comput. Vision*, 129(12), dec 2021. 2

[19] Amin Jourabloo and Xiaoming Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *CVPR*, June 2016. 2

[20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018. 3

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, pages 8110–8119, 2020. 4

[22] Kostiantyn Khabarlak and Larysa Koriashkina. Fast facial landmark detection and applications: A survey. *Journal of Computer Science and Technology*, 22(1), 2022. 2

[23] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *CVPRW*, July 2017. 2

[24] Martin Köstinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *ICCV Workshops*, pages 2144–2151, 2011. 2

[25] Abhinav Kumar, Tim K. Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *CVPR*, 2020. 6

[26] Hui Li, Zidong Guo, Seon-Min Rhee, Seungju Han, and Jae-Joon Han. Towards accurate facial landmark detection via cascaded transformers. In *CVPR*, pages 4176–4185, June 2022. 3

[27] Jinpeng Li, Haibo Jin, Shengcai Liao, Ling Shao, and Pheng-Ann Heng. Repformer: Refinement pyramid transformer for robust facial landmark detection. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*. ijcai.org, 2022. 3

[28] Weijian Li, Yuhang Lu, Kang Zheng, Haofu Liao, Chihung Lin, Jiebo Luo, Chi-Tung Cheng, Jing Xiao, Le Lu, Chang-Fu Kuo, and Shun Miao. Structured landmark detection via topology-adapting deep graph learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, 2020. 2

[29] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, pages 1954–1963, 2021. 3

[30] Yaojie Liu, Amin Jourabloo, William Ren, and Xiaoming Liu. Dense face alignment. In *ICCV Workshops*, pages 1619–1628, 2017. 2

[31] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11966–11976, 2022. 4, 6

[32] Yanda Meng, Xu Chen, Dongxu Gao, Yitian Zhao, Xiaoyun Yang, Yihong Qiao, Xiaowei Huang, and Yalin Zheng. 3d dense face alignment with fused features by aggregating cnns and gcns. *CoRR*, abs/2203.04643, 2022. 2

[33] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV Workshops*, pages 397–403, 2013. 1, 2, 4, 5, 6, 8

[34] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, October 2019. 2

[35] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, June 2020. 2

[36] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *ICCV*, Oct 2017. 2

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, page 6000–6010, 2017. 4

[38] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2019. 2

[39] Nannan Wang, Xinbo Gao, Dacheng Tao, and Xuelong Li. Facial feature point detection: A comprehensive survey. *CoRR*, abs/1410.1037, 2014. 2

[40] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *ICCV*, October 2019. 6

[41] Ukrit Watchareeruetai, Benjaphan Sommana, Sanjana Jain, Pavit Noinongyao, Ankush Ganguly, Aubin Samacoïts, Samuel W. F. Earp, and Nakarin Sritrakool. LOTR: face landmark localization using localization transformer. *IEEE Access*, 10, 2022. 3

[42] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Sebastian Dziadzio, Matthew Johnson, Virginia Estellers, Tom Cashman, and Jamie Shotton. Fake it till you make it: Face analysis in the wild using synthetic data alone. In *ICCV*, 2021. 2, 4, 5, 6

[43] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljevic, Daniel Wilde, Stephan Garbin, Chirag Raman, Jamie Shotton, Toby Sharp, Ivan Stojiljkovic, Tom Cashman, and Julien Valentin. 3d face reconstruction with dense landmarks. In *ECCV*, 2022. 1, 2, 3, 4, 6

[44] Chenglei Wu, Derek Bradley, Pablo Garrido, Michael Zollhöfer, Christian Theobalt, Markus Gross, and Thabo Beeler. Model-based teeth reconstruction. *ACM TOG*, 35(6), 2016. 4, 5

[45] Chenglei Wu, Derek Bradley, Markus Gross, and Thabo Beeler. An anatomically-constrained local deformation model for monocular face capture. *ACM TOG*, 35(4), 2016. 4, 5, 7

[46] Wenyan Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, pages 2129–2138, 2018. 2

[47] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, 2018. 6

[48] Yue Wu and Qiang Ji. Facial landmark detection: A literature survey. *Int. J. Comput. Vision*, 127(2):115–142, 2019. 2

[49] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539, 2013. 2

[50] Zixuan Xu, Banghuai Li, Ye Yuan, and Miao Geng. Anchorface: An anchor-based facial landmark detector across large poses. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3092–3100, May 2021. 2

[51] Huilin Yang, Junyan Lyu, Pujin Cheng, and Xiaoying Tang. Lddmm-face: Large deformation diffeomorphic metric learning for flexible and consistent face alignment, 2021. 2

[52] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *CVPR*, pages 598–607, 2020. 4

[53] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *CVPR*, 2021. 4, 5

[54] Yang Zhao, Yifan Liu, Chunhua Shen, Yongsheng Gao, and Shengwu Xiong. MobileFAN: Transferring deep hidden representation for face alignment. *Pattern Recognition*, 2020. 2

[55] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *ICCV Workshops*, pages 386–391, 2013. 2

[56] Meilu Zhu, Daming Shi, Mingjie Zheng, and Muhammad Sadiq. Robust facial landmark detection via occlusion-adaptive deep networks. In *CVPR*, 2019. 6

[57] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3d solution. In *CVPR*, June 2016. 2

[58] Gaspard Zoss, Thabo Beeler, Markus Gross, and Derek Bradley. Accurate markerless jaw tracking for facial performance capture. *ACM TOG*, 38(4), 2019. 4, 5