LDDMM-FACE: LARGE DEFORMATION DIFFEOMORPHIC METRIC LEARNING FOR CROSS-ANNOTATION FACE ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose an innovative, flexible and consistent cross-annotation face alignment framework, LDDMM-Face, the key contribution of which is a deformation layer that naturally embeds facial geometry in a diffeomorphic way. Instead of predicting facial landmarks via a heatmap or coordinate regression, we formulate the face alignment task in a diffeomorphic registration manner and predict momenta that uniquely parameterize the deformation between the initial boundary and true boundary. We then perform large deformation diffeomorphic metric mapping (LDDMM) simultaneously for curve and landmark to localize the facial landmarks. The novel embedding of LDDMM into a deep network allows LDDMM-Face to consistently annotate facial landmarks without ambiguity and flexibly handle various annotation schemes, and can even predict dense annotations from sparse ones. Our method can be easily integrated into various face alignment networks. We extensively evaluate LDDMM-Face on four benchmark datasets: 300W, WFLW, HELEN and COFW-68. LDDMM-Face distinguishes itself with outstanding performance when dealing with within-dataset cross-annotation learning (sparse-to-dense) and cross-dataset learning (different training and testing datasets). In addition, LDDMM-Face shows promising results on the most challenging task of cross-dataset cross-annotation learning (different training and testing datasets with different annotations).

1 INTRODUCTION

Face alignment refers to identifying the geometric structure of a human face in a digital image, through localizing key landmarks that are usually predefined and characteristic of the face's geometry. It is a prerequisite in many computer vision tasks and different numbers of facial landmarks are employed for different tasks. For example, Yi et al. (2013) performs face recognition with 34 landmarks, Yang et al. (2018) performs facial expression recognition with 63 landmarks, Taigman et al. (2014) conducts face verification with 67 landmarks, Kang et al. (2021) conducts face frontalization with 68 landmarks, and Nirkin et al. (2019) conducts face reenactment with 70 landmarks.

To accommodate different tasks, a variety of datasets with different face alignment annotation schemes have been created. For example, COFW (Burgos-Artizzu et al., 2013) annotates 29 landmarks, 300W (Sagonas et al., 2013a) annotates 68 landmarks, WFLW (Wu et al., 2018) annotates 98 landmarks, and HELEN (Le et al., 2012) annotates 194 landmarks. Most existing face alignment methods can only deal with the specific annotation scheme adopted by the training dataset, but cannot flexibly accommodate multiple annotation schemes. Therefore, if a model is trained on a dataset with a specific annotation scheme, it can then only predict landmarks of the specific scheme; a model trained on 300W with a 68-landmark annotation scheme can only predict the learned 68 landmarks but not other annotation schemes such as a 194-landmark scheme. There is no doubt that cross-annotation prediction is one of the most challenging tasks in face alignment.

In this work, we make use of facial boundaries given that they represent facial geometry well and that most facial landmarks sit on those boundaries (Wu et al., 2018). We formulate face alignment into a diffeomorphic registration framework. Specifically, we use boundary curves to represent facial geometry (Dupuis et al., 1998). Then, large deformation diffeomorphic metric mapping (LDDMM)

simultaneously for curve and landmark (Joshi & Miller, 2000; Glaunès et al., 2008) between an initial face and the true face is encoded into a neural network for landmark localization. LDDMM delivers a non-linear smooth transformation with a favorable topology-preserving one-to-one mapping property. Once the diffeomorphism, which is parameterized by momenta (Dupuis et al., 1998; Joshi & Miller, 2000; Glaunès et al., 2008), between the initial face and the true face is obtained, all points on/around the initial face have unique correspondence on/around the true face through the acquired diffeomorphism. This property makes it possible to predict facial landmarks of different annotation schemes with a model trained only on landmarks from a single annotation scheme. Utilizing both landmark and curve enables LDDMM to handle shape deformations both locally and globally for each facial boundary; the role of the landmark term is to match the corresponding landmarks whereas the role of the curve term is to make the corresponding facial curves be close to each other and to preserve facial topology. Notably, we predict momenta instead of increments between the initial face and the true face, which affords additional flexibility and is one of the key novelties of the proposed method. This is the first time that face alignment is formulated as a diffeomorphic registration problem, enabling cross-annotation as well as cross-dataset alignments.

Our main contributions are three-fold:

- We propose a novel face alignment network by integrating LDDMM into deep neural networks to handle various facial annotation schemes. Our proposed approach, LDDMM-Face, can be easily integrated into most face alignment networks to effectively predict facial landmarks with different annotation schemes.
- Our approach provides two-fold cross-annotation face alignment. 1) within-dataset cross-annotation (sparse-to-dense): training and testing on the same dataset but with different annotation schemes; 2) cross-dataset cross-annotation: training and testing on different datasets with different annotation schemes.
- We demonstrate the effectiveness of LDDMM-Face in handling challenging cases across datasets, making within-dataset cross-annotation predictions, predicting consistent facial boundaries with different training annotations, handling cross-dataset cross-annotation predictions, and accommodating various deep network settings.

2 Related Works

Depending on whether an initial face is needed as an input, existing methods can be categorized into registration-style and nonregistration-style models. Our proposed LDDMM-Face fits in the registration-style scope.

Registration-style models Registration-style models require an initial guess (usually the mean face of the training set) to serve as an input template, and then a set of parameters that characterize the deformation from the template to a target (the true face of the facial image of interest) are estimated and used to predict facial landmarks through a trained model. The set of parameters can be of various formats, including transformation parameters, displacements, and instantiating variables of deformable models. A variety of methods fall into this category, such as active shape analysis (Milborrow & Nicolls, 2008), active appearance analysis (Cootes et al., 2001; Matthews & Baker, 2004; Kahraman et al., 2007; Saragih & Goecke, 2007), supervised descent methods (Xiong & De la Torre, 2013) and others (Kazemi & Sullivan, 2014; Ren et al., 2014; Chen et al., 2014; Lee et al., 2015; Zhu et al., 2015; Tuzel et al., 2016; Su & Geng, 2019). However, none of them are based on diffeomorphic mapping which is a special class of registration delivering flexible and topology-preserving deformations. Consequently, all these aforementioned methods can only perform within-annotation prediction but not cross-annotation prediction due to their relatively constrained registration.

Cross-dataset face alignment Cross-dataset face alignment refers to training on one dataset and testing on other datasets. It can evaluate the generalization ability of a face alignment method of interest. Different datasets are usually annotated with different schemes, and thus it requires manual re-annotation to make cross-dataset face alignment plausible. For example, COFW is re-annotated with 68 landmarks (originally annotated with 29 landmarks) to perform cross-dataset evaluation in (Ghiasi & Fowlkes, 2014) and this re-annotation has also been used by many other lately-developed cross-dataset face alignment methods (Wu et al., 2018; Chen et al., 2019; Kumar et al., 2020). However, no work has ever performed cross-dataset face alignment without re-annotation.

Cross-annotation face alignment Cross-annotation face alignment refers to training on a dataset of a specific annotation scheme while testing on datasets of different annotation schemes. Cross-annotation face alignment not only measures a method's generalization ability, but also makes the task of face alignment flexible. So far, existing works (Zhu et al., 2014; Zhang et al., 2015; Wu et al., 2018) typically utilize information from multiple datasets and their corresponding annotation schemes to boost the training performance or generate pseudo-annotation on one specific dataset. No work has ever really investigated cross-annotation face alignment in terms of training on one specific annotation scheme and testing on another annotation scheme within/across datasets.

Learning-based diffeomorphic mapping Traditional diffeomorphic mapping usually requires a template-and-target pair as input for registration through minimizing their discrepancy and a regularization term for the diffeomorphic property. However, it is a one-to-one optimization process and only one mapping is obtained without utilizing any information from other available objects. In addition, it is usually time-consuming when the objects to be registered consist of numerous elements, such as 3D images or surfaces. Recently, learning-based diffeomorphic mapping through deep neural networks has been proposed in several works (Balakrishnan et al., 2018; 2019; Dalca et al., 2019), which can efficiently predict a set of mappings between the template and various targets after one-time training. However, existing learning-based diffeomorphic mapping methods mainly focus on image registration, and both template and targets are needed as the input. In face alignment, the target (true face) is not available in the testing phase, and thus existing learning-based diffeomorphic mapping methods cannot be directly applied.

LDDMM is a state-of-the-art (SOTA) diffeomorphic registration framework that has been widely used in the biomedical image field (Miller et al., 2015; Tang et al., 2015; Yang et al., 2017a; Jiang et al., 2018). Recently, LDDMM has also shown its effectiveness in facial recognition related fields (Yang et al., 2018) though under its traditional setting. LDDMM itself cannot be used to perform face alignment because it needs a pair of faces (mean face and target face) as the input. However, we can take advantage of the diffeomorphic property of LDDMM by embedding in our proposed deformation layer in LDDMM-Face, which makes it feasible to flexibly and consistently predict additional landmarks (in addition to the training ones), make cross-annotation predictions, as well as effectively deal with challenging cases.

3 Deep LDDMM Network

In LDDMM-Face, we novelly integrate LDDMM based facial shape registration into deep learning, which can consistently predict facial landmarks cross different annotations within/across datasets.

Given a normalized RGB face image, LDDMM-Face first extracts both spatial and semantic features from the input image with a replaceable backbone model. Second, the features are passed through a deep LDDMM head which consists of a momentum estimator and a deformation layer. The momentum estimator contains fully-connected layers and predicts vertical and horizontal momenta for each landmark. Suppose the geometry of a face is characterized by N boundary curves, the deformation layer has N sublayers (flow 1 to flow N). Each sublayer separately deforms the corresponding initial curve, the procedure of which is detailed in subsection 3.1. Two inputs, the mean face serving as the initial face and the estimated momenta, are fed into the deformation layer. The deformed facial curves from each layer are sequentially concatenated, yielding an estimate of the true face. Figure 1 shows the overall pipeline of LDDMM-Face.

The structure and configurations of the backbone models are identical to a SOTA facial landmark detector (Wang et al., 2020). Detailed investigations of the baseline network are not within the scope of this work. Instead, we focus on the deformation layer and the loss function since they can be readily integrated into most deep learning-based face alignment pipelines.

3.1 LDDMM Deformation Layer

3.1.1 LDDMM-curve&landmark

Our proposed deformation layer, based on LDDMM-curve&landmark, combines the advantages of LDDMM-curve (Glaunès et al., 2008) and LDDMM-landmark (Joshi & Miller, 2000) to account for both global and local discrepancies in the matching process. LDDMM (Dupuis et al., 1998; Joshi

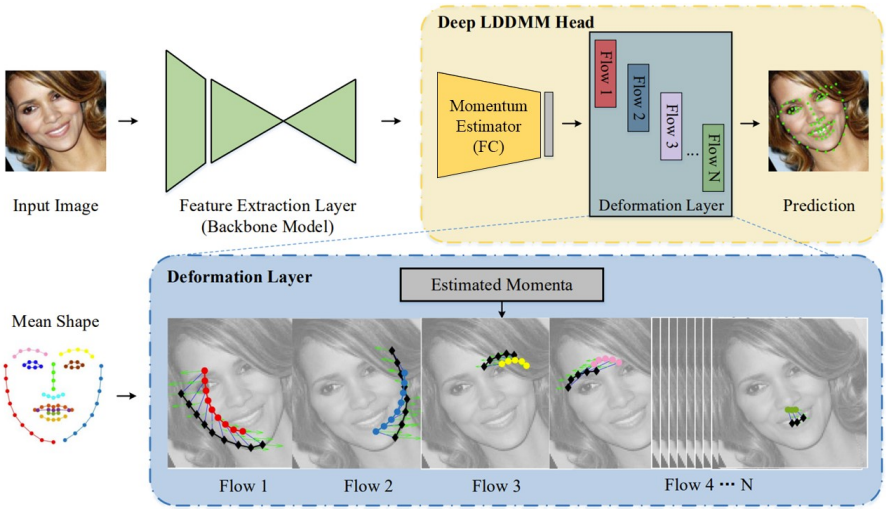

Figure 1: The overall pipeline of LDDMM-Face, which consists of a backbone model and two functional layers: a momentum estimator and a deformation layer consisting of N flows. In each flow of the deformation layer, the initial curve is shown in the same color as that in the mean face, and the deformed curve is shown in black connected diamonds. The fine blue lines connecting each initial landmark and the corresponding deformed landmark denote the trajectory of the initial landmarks. Green arrows show the predicted momenta at each time step along the trajectory.

& Miller, 2000; Glaunès et al., 2008) is a registration framework that provides a diffeomorphic transformation acting on the ambient space. Under the LDDMM framework, objects are placed into a reproducing kernel Hilbert metric space through time-varying velocity vector fields $v_t(\cdot)$: $\mathbb{R}^2 \rightarrow \mathbb{R}^2$ for $t \in [0, 1]$ in the ambient space. The underlying assumption is that the two objects of interest are of equivalent topology and one can be deformed to the other via a flow of diffeomorphisms. Given a pair of objects C and S, the time-dependent flow of diffeomorphisms transforming C to S is defined according to the ordinary differential equation (ODE) $\dot{\phi}_t(x) = v_t(\phi_t(x))$, with $\phi_0(x) = x$. The resulting diffeomorphism $\phi_1(x)$ is acquired as the end point of the diffeomorphism flow at time $t = 1$ such that $\phi_1 \cdot C = S$. To ensure the resulting transformation is diffeomorphic, v_t must satisfy the constraint that $\int_0^1 ||v_t||_V dt < \infty$, with V being a Hilbert space associated with a reproducing kernel function k_V and a norm $\| \cdot \|_V$ (Trouvé, 1995). In practice, a Gaussian kernel is selected for k_V so that $k_V(a, b) = exp(-\frac{\|a-b\|_2^2}{\sigma_V^2})$, wherein σ_V represents the kernel size that is usually empirically selected and $\| \ \|_2$ denotes the l^2-norm.

In LDDMM-curve, a curve C_c is discretized into a sequence of n ordered points $\mathbf{x} = (x_i)_{i=1}^n$. The curve can be encoded by those points along with their tangent vectors such that $C_c = (c_{\mathbf{x},i}, \tau_{\mathbf{x},i})_{i=1}^n$, with $c_{\mathbf{x},i} = \frac{x_{i+1}+x_i}{2}$ being the center of two sequential points and $\tau_{\mathbf{x},i} = x_{i+1} - x_i$ being the tangent vector at point $c_{\mathbf{x},i}$. C_c is associated with a sum of vector-valued Diracs, $\mu_{C_c} = \sum_{i=1}^n \tau_{\mathbf{x},i} \delta_{c_{\mathbf{x},i}}$, and is embedded into a Hilbert metric space W of smooth vectors with the norm being

$$\begin{aligned} \|\mu_{C_c}\|_{W^*}^2 &= \left\| \sum_{i=1}^n \tau_{\mathbf{x},i} \delta_{c_{\mathbf{x},i}} \right\|_{W^*}^2 \\ &= \sum_{i=1}^n \sum_{j=1}^n k_W(c_{\mathbf{x},i}, c_{\mathbf{x},j}) \tau_{\mathbf{x},i} \cdot \tau_{\mathbf{x},j}, \end{aligned} \tag{1}$$

where k_W is the reproducing kernel in the space W (k_W is of the same form as that of k_V) and W^* is the dual space of W. In LDDMM-landmark, a set of n landmarks C_l are represented by its Cartesian coordinates. Thus, a set of ordered points can be modelled as both curve and landmark.

LDDMM-curve can handle the overall shape whereas LDDMM-landmark is more powerful in dealing with local information. Assume that the template object C (the transforming object) and the

target object S (the object being transformed to) are respectively discretized as $\mathbf{x} = (x_i)_{i=1}^n$ and $\mathbf{y} = (y_i)_{i=1}^n$, and $\mathbf{z} = (z_i)_{i=1}^n$ is the deformed object $\phi_1 \cdot C$, then the resulting diffeomorphism ϕ_1 is obtained by minimizing the following inexact matching functional

$$J_{c,s}(v_t) = \min_{v_t:\dot{\phi}_t = v_t(\phi_t), \phi_0 = id} \gamma \int_0^1 ||v_t||_V^2 dt + D(\phi_1 \cdot C, S), \tag{2}$$

where $\int_0^1 ||v_t||_V^2 dt$ can be interpreted as the energy consumed by the flow of diffeomorphisms, and the second term quantifies the overall discrepancy between the deformed object $\phi_1 \cdot C$ and the target object S. γ is a weight in $[0, 1]$ serving as the trade-off coefficient between the consumed energy and the overall discrepancy. In LDDMM-curve&landmark, the discrepancy consists of two parts

$$D(\phi_1 \cdot C, S) = \beta D_c(\phi_1 \cdot C_c, S_c) + D_l(\phi_1 \cdot C_l, S_l), \tag{3}$$

where D_c measures the discrepancy between the deformed object and the target object when modelled as curves and D_l quantifies the corresponding discrepancy when modelled as landmarks. β is a trade-off weight deciding the relative importance of curve and landmark. The curve discrepancy is computed as the norm of the difference between the two vector-valued curve representations in the space W^*, which is explicitly

$$D_c(\phi_1 \cdot C_c, S_c) = || \sum_{i=1}^n \tau_{\mathbf{z},i} \delta_{c_{\mathbf{z},i}} - \sum_{i=1}^n \tau_{\mathbf{y},j} \delta_{c_{\mathbf{y},j}} ||_{W^*}^2, \tag{4}$$

and the landmark discrepancy is computed as the Euclidean distance averaged across all point pairs

$$D_l(\phi_1 \cdot C_l, S_l) = \frac{1}{n} \sum_{i=1}^n \|\mathbf{z}_i - \mathbf{y}_i\|_2 . \tag{5}$$

After minimizing $J_{c,s}(v_t)$, the resulting diffeomorphism ϕ_1 is parameterized by the velocity vector field $v_t(x)$ as $v_t(x) = \sum_{i=1}^n k_V(x_i(t), x)\alpha_i(t)$, where $\alpha_i(t)$ denotes the time-dependent momentum at the i-th landmark. A diffeomorphism is completely encoded by the initial momenta in the template space. These momenta can be obtained by solving the following sets of ODEs

$$\frac{dx_i(t)}{dt} = \sum_{j=1}^n k_V(x_j(t), x_i(t))\alpha_j(t), i = 1, \ldots, n, \tag{6}$$

where $x_i(t), t \in [0, 1]$ denotes the trajectory of the i-th landmark on the template object.

3.1.2 Deformation Layer

The deformation layer takes the predicted momenta as inputs to perform LDDMM-induced deformation on the initial (mean) face. Trajectory of the i-th landmark follows

$$x_i(t) = x_i(0) + \int_0^1 (\sum_{j=1}^n k_V(x_j(t), x_i(t))\alpha_j(t))dt. \tag{7}$$

The finally estimated true face (also called deformed face) is obtained at the end time point of the transformation flow.

As illustrated in the top panel of Figure 1, since a face is modelled using N boundary curves, the LDDMM transformation component of the deformation layer is implemented separately for each curve from flow 1 to flow N. N depends on the annotation scheme. The procedure of each flow is demonstrated in the bottom panel of Figure 1.

3.2 Loss Function

The loss function in our proposed network is inspired by the objective function of LDDMM-curve&landmark in Eq. 2. Focusing on accuracy, γ is chosen to be 0 given that an accurate matching matters more than a geodesic path in face alignment. Although γ is 0, the solution of the loss function is embedded into the V space and still yields diffeomorphic transformations. Therefore, the loss function, minimized with respect to the momenta α in LDDMM is

$$\min_{\alpha} \frac{\sum_{p=1}^N \beta D_{pc}(S_{cp}^{\text{deform}}, S_{cp}^*) + D_{pl}(S_{lp}^{\text{deform}}, S_{lp}^*)}{d_{ipd}}, \tag{8}$$

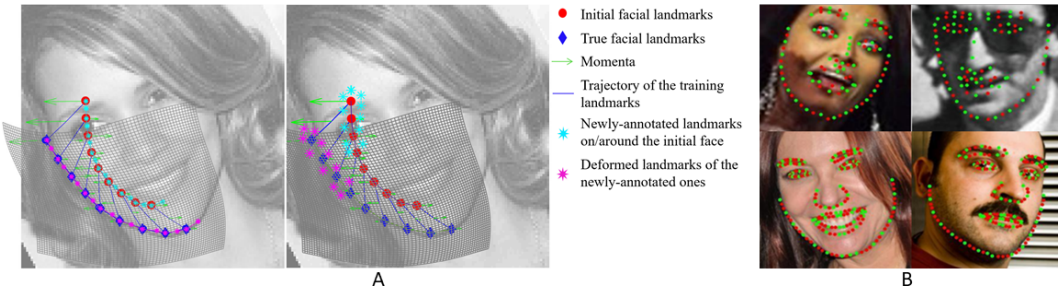

Figure 2: Demonstration of flexible and consistent face alignment for a right cheek (A), and cross-annotation face alignment (sparse-to-dense) results on WFLW and HELEN (B). In A, red circles and blue diamonds respectively represent the initial and true facial landmarks used in training. Cyan stars and magenta stars respectively represent newly-annotated landmarks on/around the initial face and the corresponding deformed landmarks of the newly-annotated ones through diffeomorphism obtained in the training stage. Green arrows denote the momenta along the trajectory of the transforming curve and blue lines represent the corresponding trajectory. Gray grids represent the diffeomorphism-induced deformations. Note that the cyan points not used in computing the diffeomorphism are still deformed correctly. In B, green dots represent landmarks involved in training and red dots represent additionally predicted landmarks. Top row indicates results on WFLW (50% for training) and bottom row indicates results on HELEN (33% for training).

where S^*_{cp} is the vector-measured expression of the ground truth curve of the p-th facial curve, S^{deform}_{cp} denotes the corresponding deformed curve, and D_{pc} quantifies (via Eq. 4) the discrepancy between the ground truth and the deformed curve. S^*_{lp} is a vector representing the ground truth landmarks of the p-th facial curve, S^{deform}_{lp} denotes the corresponding deformed landmarks and D_{pl} measures (via Eq. 5) the discrepancy between the ground truth and the deformed landmarks. d_{ipd} is the distance between the pupils of the ground truth face. β is a trade-off coefficient between landmarks and curves. In this way, our loss function takes discrepancies of both landmarks and curves into consideration and consequentially is able to handle local as well as global discrepancies between the true face and the deformed face.

3.3 Flexible and Consistent Face Alignment

Once momenta are obtained from LDDMM between an initial face and a true face, the diffeomorphism between that face pair is uniquely defined (Joshi & Miller, 2000; Glaunès et al., 2008). This transformation can be used to deform not only landmarks used in the matching procedure but also any other landmarks sitting around the transforming face boundary. Due to the smooth, topology-preserving and one-to-one mapping property of the obtained diffeomorphism, we can compute the deformed location of any landmark lying on/around the face boundary in a consistent way. Any two deformed landmarks would never come across each other and any deformed boundary would never cross itself, which is practical and intuitive for muscle motion of the human face. Suppose the initial locations of m landmarks lying on/around the face boundary are $a(0)$, we have $a_k(t) = a_k(0) + \int_0^1 (\sum_{j=1}^n k_V(x_j(t), a_k(t))\alpha_j(t))dt$, where $a_k(t)$ represents the location of the deformed k-th landmark at time t, $k = 1, \ldots, m$. $x_j(t)$ and $\alpha_j(t)$ respectively denote the location and momentum of the j-th landmark. k_V is the reproducing kernel. The final locations are obtained at the end point of the transformation flow, namely $a_k(1)$. The transformations and acquired momenta are different for the N facial curves, and thus the landmarks on/around each curve are deformed separately.

Figure 2 demonstrates an example of flexible and consistent alignment, in which some of the newly-annotated cyan star landmarks which were not involved in obtaining the LDDMM-induced diffeomorphism can still be deformed to proper locations through the predicted diffeomorphism.

Therefore, given a pair of initial and true faces, once we have the LDDMM derived momenta, we can flexibly as well as consistently predict the deformed location of any extra landmark regardless of the training annotation scheme used.

3.4 CROSS-ANNOTATION FACE ALIGNMENT

Due to LDDMM-Face's ability to perform flexible and consistent face alignment (see subsection 3.3), we can easily conduct cross-annotation face alignment in two different ways. The first is to perform cross-annotation face alignment within the same dataset but in a sparse-to-dense manner. The second is to perform cross-annotation face alignment across different datasets. The implemental details are described in subsections 4.2 and 4.3.

4 EXPERIMENT

In this section, the employed datasets, the error metrics, and the implementation details are described. Subsection 4.1 validates the diffeomorphic transformation of LDDMM-Face. Subsections 4.2 and 4.3 show the flexibility and consistency of LDDMM-Face by performing cross-annotation face alignment in both sparsely-supervised (spares-to-dense) and cross-dataset respects.

Datasets To evaluate the performance of LDDMM-Face, we conduct experiments on 300W (Sagonas et al., 2013a), WFLW (Wu et al., 2018), HELEN (Le et al., 2012) and COFW-68 (Burgos-Artizzu et al., 2013; Ghiasi & Fowlkes, 2015), all of which are benchmark datasets for face alignment. For more details on these datasets, please refer to Appendix A.

Error Metrics We use two main metrics to quantify the face alignment error:

- NME_l: The mean distance between the predicted landmarks and the ground truth landmarks divided by the inter-ocular distance (Ren et al., 2014; Zhu et al., 2015; Xiao et al., 2016).
- NME_c: The mean iterative closest point (ICP) error between the predicted curves and the ground truth curves divided by the inter-ocular distance (Arun et al., 1987).

Specifically, the ICP error is introduced to quantify the overall curve discrepancy and it can be used to solve the problem that inter-ocular landmark distance is unavailable when there is no point-by-point correspondence between the predicted landmarks and the ground truth landmarks. With the ICP error, fair comparisons can be conducted between the baseline method (w/o LDDMM-Face) and LDDMM-Face in cross-annotation settings. The failure rate ($\text{FR}_{0.1}$) is further used for COFW-68 to be consistent with previous studies (Zhu et al., 2015; Wu & Yang, 2017).

Implementation LDDMM-Face consists of a backbone model, a momentum estimator, a deformation layer and a loss function, as described in section 3. For the backbone model, we employ HRNet (Wang et al., 2020) for its SOTA performance on face alignment. Other backbone models are also investigated to evaluate the adaptability of LDDMM-Face (see Appendix C). For the momentum estimator, we adopt a simple yet effective structure consisting of an average pooling layer and a fully-connected layer (same structure as the coordinate regression head). For the deformation layer of LDDMM-curve&landmark, σ_V and σ_W are respectively chosen to be the scale and half the scale of the coordinates of each curve of the mean face. N is chosen to be 12 in order to efficiently characterize different parts of a face. For the loss function, β is empirically chosen to be 0.1. (see Table 5 in Appendix B). All experiments are conducted with PyTorch 1.7.1 (Paszke et al., 2019) on 4 RTX 3090 GPUs. More details are provided in Appendix B.

4.1 DIFFEOMORHPIC FACE ALIGNMENT

We first validate the predicted mapping's smoothness (diffeomorphic property) even after removing the energy term ($\gamma = 0$). Following the criteria of diffeomorphism (Balakrishnan et al., 2019), we compute the determinant of Jacobian (DetJ) of the predicted diffeomorphisms for all landmarks and the averaged value (DetJ_A). We also compute the total number of landmarks whose determinants of Jacobian are less than or equal to zero (DetJ_N) and the determinant of Jacobian (DetJ_AM) of the averaged predicted diffeomorphisms for all landmarks. The corresponding results on 300W and WFLW are shown in Table 1. Both DetJ_A and DetJ_AM are larger than zero, and there is no landmark whose DetJ is less than or equal to zero. All these results demonstrate the diffeomorphism property of the transformations obtained from LDDMM-Face.

Dataset	DetJ_A	DetJ_N	DetJ_AM
300W	1.0416	0	1.0102
WFLW	1.0001	0	0.9990

Table 1: The Jacobian determinant analysis results of LDDMM-Face.

4.2 Cross-annotation Face Alignment in a Sparse-to-dense Manner

We validate the flexibility and consistency of LDDMM-Face by performing sparse-to-dense cross-annotation face alignment within various datasets. As described in subsection 3.3, LDDMM-Face can predict any extra landmark lying nearby a predefined curve. Therefore, we can train a model with sparse landmarks (a subset of the full annotations) that minimally describe facial geometry to predict dense landmarks in a consistent way. By dense, we mean the full annotations. We conduct such experiments on 300W, WFLW and HELEN. For 300W and WFLW, 50% facial landmarks are used for sparsely-supervised training. Since HELEN has a total of 194 annotated landmarks, which is a relatively large number, we further reduce the training landmarks to 33% in the HELEN experiment. As tabulated in Table 2, LDDMM-Face outperforms its baseline by a large margin. In terms of NME_c, there is a 40% improvement on 300W, a 10% improvement on WFLW and a 20% improvement on HELEN, when trained with 50% landmarks. A 35% improvement is observed on HELEN when trained with 33% landmarks. Notably, LDDMM-Face is much better than the baseline in detecting face contour and eyebrows, indicating it works better for curves with large deformations. When trained on sparse landmarks, there is only very mild decline in LDDMM-Face's performance compared to training on full landmarks, and it works even better than some of the fully-supervised methods. Other methods cannot perform such sparse-to-dense predictions. In Figure 3.3 we show some representative qualitative results and more in Appendix D.

Methods	Dataset	TL	NME_c (%)						NME_l (%)
			O	F	E	N	I	M	
HRNet	300W	50%	4.82	9.34	6.00	3.18	**1.86**	3.74	-
w. LDDMM-Face			**2.94**	**4.82**	**3.47**	**2.28**	1.87	**2.25**	**3.18**
HRNet	HELEN	50%	2.95	4.33	3.08	3.16	1.77	2.40	-
w. LDDMM-Face			**2.39**	**3.37**	**2.52**	**2.61**	**1.46**	**1.97**	**3.71**
HRNet	HELEN	33%	3.73	5.63	3.95	3.92	2.15	3.01	-
w. LDDMM-Face			**2.45**	**3.29**	**2.74**	**2.76**	**1.56**	**1.91**	**3.78**
HRNet	WFLW	50%	3.95	5.72	4.04	3.34	3.30	**3.36**	-
w. LDDMM-Face			**3.58**	**4.67**	**3.72**	**3.20**	**2.95**	3.38	**4.79**

Table 2: Sparse-to-dense prediction results on the 300W common set, HELEN test set and WFLW test set. 'TL' means the fraction of full landmarks used during training. 'O' indicates overall face. 'F' means face contour. 'E' means eyebrows. 'N' means nose. 'I' means eyes. 'M' means mouth. '-' indicates NME_l is unavailable for HRNet since it cannot make cross-annotation predictions.

4.3 Cross-annotation Face Alignment across Datasets

We further validate the flexibility and consistency of LDDMM-Face by evaluating the performance of cross-annotation face alignment across different datasets. Existing cross-dataset evaluations mainly utilize the COFW-68 dataset which has been re-annotated with an identical scheme as that of 300W. Cross-annotation is not feasible for existing face alignment methods.

As mentioned above, HELEN has two annotation schemes since it is also a subset of 300W. As such, the cross-annotation experiment between HELEN and 300W can be seen as cross-annotation but within-dataset. By conducting an affine transformation from source mean face to target mean face, we can easily predict landmarks of different annotation schemes without retraining. From Table 3, we observe that LDDMM-Face significantly improves the performance over the baseline. It should be noted that although the 194-landmark annotation scheme of HELEN describes the nose and eyebrow in totally different ways from the 68-landmark annotation scheme of 300W, LDDMM-Face achieves decent results. We also conduct simultaneous cross-dataset and cross-annotation experiments between 300W and WFLW, on which only slight improvements are observed due to the highly similar annotation schemes between these two datasets. Table 3 shows that LDDMM-Face

Methods	Train set	Test set	NME_c (%)						NME_l (%)
			O	F	E	N	I	M	
HRNet	HELEN	300W	5.49	**6.72**	5.82	9.00	2.72	3.21	-
w. LDDMM-Face			**4.76**	6.81	**4.25**	**7.45**	**2.21**	**3.09**	**5.96**
HRNet	300W	HELEN	5.60	6.61	6.18	9.07	2.79	3.36	-
w. LDDMM-Face			**4.13**	**3.47**	**4.79**	**7.19**	**2.41**	**2.81**	**7.58**
HRNet	WFLW	300W	3.91	5.29	4.62	**3.01**	3.14	**3.46**	-
w. LDDMM-Face			**3.88**	**5.27**	**4.08**	3.32	**2.94**	3.76	**4.53**
HRNet	300W	WFLW	6.61	8.65	6.74	**5.63**	6.02	6.02	-
w. LDDMM-Face			**6.04**	**6.82**	**6.41**	5.77	**5.33**	**5.88**	**9.58**

Table 3: Comparisons between LDDMM-Face and HRNet on the 300W common set, HELEN test set and WFLW test set for cross-dataset/annotation face alignment.

Method	NME_l(%)	$FR_{0.1}$(%)
TCDCN (Zhang et al., 2016)	7.66	16.17
SAPM (Ghiasi et al., 2015)	6.64	5.72
CFSS (Zhu et al., 2015)	6.28	9.07
HRNet (Wang et al., 2020)	4.97	3.16
Softlabel (Pretrained) (Chen et al., 2019)	4.82	-
LAB (Extra Data) (Wu et al., 2018)	4.62	2.17
LUVLi (Kumar et al., 2020)	**4.54**	-
LDDMM-Face	**4.54**	**1.18**

Table 4: NME_l and $FR_{0.1}$ results of training on 300W and testing on the COFW-68 test set.

is much better than the baseline in NME_c, but NME_l is still relatively unsatisfactory compared to traditional within-dataset within-annotation predictions. A plausible reason is that we use an affine transformation between the two different mean faces rather than directly modify the mean face used in the specific training process, and the two mean faces may be highly inconsistent with each other. With that being said, this is to the best of our knowledge the first attempt of simultaneous cross-dataset and cross-annotation face alignment, with satisfactory performance in identifying the overall facial geometry (curve error). This observation further verifies the effectiveness and importance of LDDMM-Face.

To compare with existing SOTA cross-dataset face alignment results, we further conduct experiments on COFW-68, the results of which are tabulated in Table 4. LDDMM-Face significantly outperforms those methods being compared, especially in terms of $FR_{0.1}$ which is very sensitive to challenging cases like large pose and occlusion. The superior performance of LDDMM-Face for challenging cases is mainly due to the curve and landmark induced diffeomorphism; diffeomorphic transforming ensures the deformed facial geometry is consistent with that of the initial face such that the occluded parts can still be accurately predicted. Collectively, LDDMM-Face makes precise facial geometry predictions across different annotations (both within and across datasets) and performs outstandingly for cross-dataset settings. More cross-dataset/annotation results can be found in Appendix E.

5 Conclusion

In this work, we present and validate a novel face alignment pipeline, LDDMM-Face, that is able to perform cross-annotation face alignment in both sparsely-supervised and cross-dataset manners. The flexibility and consistency delivered by LDDMM-Face arise naturally from an embedding of LDDMM into deep learning. It bridges the gap between different annotation schemes and makes the task of face alignment more flexible than existing methods which can only predict landmarks used in annotations of the training data. Furthermore, LDDMM-Face generalizes well and can be integrated into various deep learning based face alignment networks.

REFERENCES

K Somani Arun, Thomas S Huang, and Steven D Blostein. Least-squares fitting of two 3-d point sets. *IEEE Transactions on pattern analysis and machine intelligence*, (5):698–700, 1987.

Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9252–9260, 2018.

Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.

Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940, 2013.

Bjorn Browatzki and Christian Wallraven. 3fabrec: Fast few-shot face alignment by reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6110–6120, 2020.

Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1513–1520, 2013.

Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In *European Conference on Computer Vision*, pp. 109–122. Springer, 2014.

Lisha Chen, Hui Su, and Qiang Ji. Face alignment with kernel density deep neural network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6992–7002, 2019.

Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):681–685, 2001.

Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical image analysis*, 57: 226–236, 2019.

Arnaud Dapogny, Kevin Bailly, and Matthieu Cord. Decafa: deep convolutional cascade for face alignment in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6893–6901, 2019.

Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 379–388, 2018.

Paul Dupuis, Ulf Grenander, and Michael I Miller. Variational problems on flows of diffeomorphisms for image matching. *Quarterly of applied mathematics*, pp. 587–600, 1998.

Golnaz Ghiasi and Charless C Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2385–2392, 2014.

Golnaz Ghiasi and Charless C Fowlkes. Occlusion coherence: Detecting and localizing occluded faces. *arXiv preprint arXiv:1506.08347*, 2015.

Golnaz Ghiasi, Charless C Fowlkes, and C Irvine. Using segmentation to predict the absence of occluded parts. In *BMVC*, pp. 22–1, 2015.

Joan Glaunès, Anqi Qiu, Michael I Miller, and Laurent Younes. Large deformation diffeomorphic metric curve mapping. *International journal of computer vision*, 80(3):317, 2008.

Zihan Jiang, Huilin Yang, and Xiaoying Tang. Deformation-based statistical shape analysis of the corpus callosum in mild cognitive impairment and alzheimer's disease. *Current Alzheimer Research*, 15(12):1151–1160, 2018.

Sarang C Joshi and Michael I Miller. Landmark matching via large deformation diffeomorphisms. *IEEE transactions on image processing*, 9(8):1357–1370, 2000.

Fatih Kahraman, Muhittin Gokmen, Sune Darkner, and Rasmus Larsen. An active illumination and appearance (aia) model for face alignment. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–7. IEEE, 2007.

Zhiqi Kang, Radu Horaud, and Mostafa Sadeghi. Robust face frontalization for visual speech recognition. In *International Conference on Computer Vision Workshops*, 2021.

Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1867–1874, 2014.

Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 88–97, 2017.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8236–8246, 2020.

Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *European conference on computer vision*, pp. 679–692. Springer, 2012.

Donghoon Lee, Hyunsin Park, and Chang D Yoo. Face alignment using cascade gaussian process regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4204–4212, 2015.

Peter M. Roth Martin Koestinger, Paul Wohlhart and Horst Bischof. Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

Iain Matthews and Simon Baker. Active appearance models revisited. *International journal of computer vision*, 60(2):135–164, 2004.

Stephen Milborrow and Fred Nicolls. Locating facial features with an extended active shape model. In *European conference on computer vision*, pp. 504–513. Springer, 2008.

Michael I Miller, Laurent Younes, J Tilak Ratnanather, Timothy Brown, Huong Trinh, David S Lee, Daniel Tward, Pamela B Mahon, Susumu Mori, Marilyn Albert, et al. Amygdalar atrophy in symptomatic alzheimer's disease based on diffeomorphometry: the biocard cohort. *Neurobiology of aging*, 36:S3–S10, 2015.

Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7184–7193, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Shengju Qian, Keqiang Sun, Wayne Wu, Chen Qian, and Jiaya Jia. Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10153–10163, 2019.

Deva Ramanan and Xiangxin Zhu. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 2879–2886. IEEE, 2012.

Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1685–1692, 2014.

Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 397–403, 2013a.

Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 896–903, 2013b.

Jason Saragih and Roland Goecke. A nonlinear discriminative approach to aam fitting. In *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8. IEEE, 2007.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Kai Su and Xin Geng. Soft facial landmark detection by label distribution learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5008–5015, 2019.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1701–1708, 2014.

Xiaoying Tang, Dominic Holland, Anders M Dale, Laurent Younes, Michael I Miller, and Alzheimer's Disease Neuroimaging Initiative. The diffeomorphometry of regional shape change rates and its relevance to cognitive deterioration in mild cognitive impairment and a lzheimer's disease. *Human brain mapping*, 36(6):2093–2117, 2015.

George Trigeorgis, Patrick Snape, Mihalis A Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4177–4187, 2016.

Alain Trouvé. An infinite dimensional group approach for physics based models in pattern recognition. *preprint*, 1995.

Oncel Tuzel, Tim K Marks, and Salil Tambe. Robust face alignment using a mixture of invariant experts. In *European Conference on Computer Vision*, pp. 825–841. Springer, 2016.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Wenyan Wu and Shuo Yang. Leveraging intra and inter-dataset variations for robust face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 150–159, 2017.

Shengtao Xiao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, and Ashraf Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *European conference on computer vision*, pp. 57–72. Springer, 2016.

Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 532–539, 2013.

Zixuan Xu, Banghuai Li, Miao Geng, and Ye Yuan. Anchorface: An anchor-based facial landmark detector across large poses. *AAAI*, 2021.

Huilin Yang, Jing Wang, Haiyun Tang, Qinle Ba, Ge Yang, and Xiaoying Tang. Analysis of mitochondrial shape dynamics using large deformation diffeomorphic metric curve matching. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4062–4065. IEEE, 2017a.

Jing Yang, Qingshan Liu, and Kaihua Zhang. Stacked hourglass network for robust facial landmark localisation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 79–87, 2017b.

Pucheng Yang, Huilin Yang, Yuanyuan Wei, and Xiaoying Tang. Geometry-based facial expression recognition via large deformation diffeomorphic metric curve mapping. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 1937–1941. IEEE, 2018.

Dong Yi, Zhen Lei, and Stan Z. Li. Towards pose robust face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Leveraging datasets with varying annotations for face alignment via deep regression network. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3801–3809, 2015.

Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):918–930, 2016.

Meilu Zhu, Daming Shi, Mingjie Zheng, and Muhammad Sadiq. Robust facial landmark detection via occlusion-adaptive deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3486–3496, 2019.

Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Transferring landmark annotations for cross-dataset face alignment. *arXiv preprint arXiv:1409.0602*, 2014.

Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4998–5006, 2015.

A DATASETS

To evaluate the performance of LDDMM-Face, we conduct experiments on 300W (Sagonas et al., 2013a), WFLW (Wu et al., 2018), HELEN (Le et al., 2012), COFW-68 (Burgos-Artizzu et al., 2013; Ghiasi & Fowlkes, 2015) and AFLW Martin Koestinger & Bischof (2011), all of which are benchmark datasets for face alignment.

300W contains five different datasets: LFPW (Belhumeur et al., 2013), AFW (Ramanan & Zhu, 2012), HELEN (Le et al., 2012), IBUG (Sagonas et al., 2013a) and 300W private test set (Sagonas et al., 2013a). Each image in this dataset is annotated with 68 landmarks (Sagonas et al., 2013b) and equipped with a bounding box generated by a face detector. Following the configurations of most existing methods, 300W is divided into training and full sets. The training set contains the training sets of LFPW, HELEN and AFW, including a total of 3148 images. The full set contains the test sets of LFPW and HELEN plus the IBUG dataset (689 images in total). For a comprehensive comparison with existing methods, the test set is further split into four subsets:

- Common subset: consisting of the test sets of LFPW and HELEN (554 images in total)
- Challenging subset: consisting of the entire IBUG dataset (135 images in total)
- Fullset/300W public test set: consisting of the test sets of LFPW and HELEN plus the IBUG dataset (689 images in total)
- 300W private test set (600 images in total)

Following (Wu et al., 2018), WFLW is also divided into a training set consisting of 7500 faces and a testing set consisting of 2500 faces, with a 98-landmark annotation scheme. It is the most challenging face alignment dataset which has large variations in expression, pose and occlusion; specifically, WFLW consists of six challenging categories, namely "Pose", "Expression", "Illumination", "Make-up", "Occlusion" and "Blur". HELEN consists of 2000 training and 330 testing facial images in the wild, annotated with 194 landmarks. Its dense annotations enable us to well validate our sparsely-supervised face alignment performance. COFW-68 is a dataset featured on multiple externally occluded landmarks, which consists of 507 testing images. COFW-68 is annotated with 68 landmarks of the same annotation scheme as that of 300W. AFLW contains about 25000 facial images which are annotated with up to 21 landmarks per image.

B IMPLEMENTATION DETAILS

Data augmentations including mirroring around the Y axis, random rotation ($\pm$ 20 degrees) and scaling ($\pm$ 25%) sampled from normal distributions are applied to each training image. All images are cropped and resized to 256$\times$256 according to the provided bounding boxes, and are normalized by subtracting the mean and then getting divided by the standard deviation of the training set for each RGB channel.

Figure 3, Figure 4 and Figure 5 respectively show our mean face for 300W, HELEN and WFLW. Please note the mean face for COFW-68 is the same as that for 300W. In our sparsely-supervised face alignment experiments, we do not remove any landmarks from the eye curve for either 300W or WFLW. This is because they are already the minimum landmarks that can sufficiently describe the geometry of the corresponding curve.

We evaluate the impact of different β in Table 5. When β is small (e.g., $\beta = 0.01$), the model will be more driven by the landmark loss, resulting in small NME_l errors. However, in this situation, the alignment performance will degrade in terms of the overall facial geometry matching, especially in the sparsely-supervised learning framework (bottom panel of Table 5) since there is only a limited number of landmarks. When β is large (e.g., $\beta = 1$), the facial landmark detection accuracy will be significantly deteriorated. In light of this, β is chosen to be 0.1.

We implement the deformation layer and the loss function efficiently. With a batch size of 32 and the same backbone model, LDDMM-Face runs significantly faster than HRNet. However, this speed efficiency comes at the cost of a little bit more GPU memory usage which grows larger when there are more landmarks. Details are shown in Table 6.

Figure 3: Mean faces of full landmarks and 50% landmarks for 300W and COFW-68.

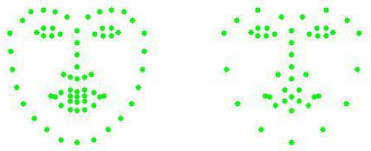

Figure 4: Mean faces of full landmarks, 50% landmarks and 33% landmarks for HELEN .

Figure 5: Mean faces of full landmarks and 50% landmarks for WFLW.

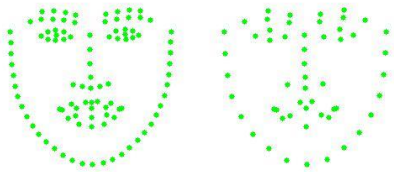

Table 5: Evaluation of the influence of different values of β on the 300W common set. 'Train' means the training fraction of full landmarks used in sparsely-supervised face alignment. Sample logic applies for 'Test'.

β	Train	Test	NME_c (%)						NME_l (%)
			O	F	E	N	I	M	
1	50%	50%	3.80	5.97	4.51	3.07	2.76	2.70	5.41
0.1			**3.41**	**5.81**	**4.09**	2.70	2.16	2.27	3.31
0.01			3.46	5.90	4.50	**2.64**	**2.07**	**2.21**	**3.31**
1	50%	100%	4.80	5.71	7.28	4.08	2.76	4.19	7.02
0.1			**3.27**	**5.32**	**3.76**	**2.63**	**2.16**	**2.50**	**3.67**
0.01			4.42	5.71	7.17	3.75	2.07	3.40	5.78

Table 6: Training speed and memory usage comparisons between HRNet and LDDMM-Face.

Method	Memory Usage (MB)	Speed (samples/s)
	50-landmark annotation	
HRNet	**6999**	45.3
LDDMM-Face	7039	**51.6**
	98-landmark annotation	
HRNet	**7056**	24.8
LDDMM-Face	7815	**34.4**
	194-landmark annotation	
HRNet	**7121**	17.0
LDDMM-Face	11449	**22.3**

We use ADAM to optimize the model with a batch size of 32 for 300 epochs. The learning rate is initialize as 3×10^{-4} and is decayed by 2 at the 75*th*, the 150*th* and the 225*th* epochs.

Codes are available at https://github.com/ForTest66656/ForTest.

C ADAPTIVE FACE ALIGNMENT ACROSS DIFFERENT LEARNING FRAMEWORKS

In this section, we investigate the adaptability and robustness of LDDMM-Face incorporated into three networks, namely HRNet (Wang et al., 2020), Hourglass (Yang et al., 2017b) and DAN (Kowalski et al., 2017). HRNet and Hourglass are SOTA heatmap regression methods and DAN is a multi-stage coordinate regression method. Considering computation cost, we use HRNetV2-W18, 2-stacked hourglass and VGG11 (Simonyan & Zisserman, 2014) as the corresponding backbone models of the three networks. Experimental results on 300W, WFLW and HELEN (Table 7) demonstrate that LDDMM-Face can be easily integrated into those face alignment networks. Among all these three settings, LDDMM-Face with HRNet gives the best results, which is the reason why we employ HRNet as our default baseline model.

Table 7: NME_l of LDDMM-Face incorporated into three different face alignment settings, obtained on the 300W full set, WFLW test set and HELEN test set.

Method	$\text{NME}_l(\%)$		
	300W	WFLW	HELEN
HRNet w. LDDMM-Face	**3.53**	**4.63**	**3.57**
Hourglass w. LDDMM-Face	3.73	5.00	3.89
DAN w. LDDMM-Face	3.91	5.43	3.95

D COMPARISON WITH STATE-OF-THE-ART RESULTS

We compare LDDMM-Face with SOTA approaches on the test sets of 300W, WFLW and AFLW, respectively in Table 8, Table 9 and Table 10. Experimental results verify the effectiveness of LDDMM-Face. For 300W, the performance of LDDMM-Face is comparable to its baseline HRNet and outperforms most existing methods. For WFLW, although this dataset confronts large variations of poses, expressions and occlusions, LDDMM-Face yields superior results and outperforms almost all compared approaches. For AFLW, LDDMM-Face shows its robustness under sparse annotation schemes.

Table 8: NME_l results on the 300W common set, challenging set and full set. "Sparse-LF" is short for training landmark fraction in sparsely-supervised face alignment.

Method	$\text{NME}_l(\%)$		
	300W Common	300W Challenging	300W Full
MDM (Trigeorgis et al., 2016)	4.83	10.14	5.88
RAR (Xiao et al., 2016)	5.03	8.95	5.80
DAN (Kowalski et al., 2017)	3.15	5.53	3.62
SAN (Dong et al., 2018)	3.34	6.60	3.98
LAB (Extra Data) (Wu et al., 2018)	2.98	5.19	3.49
ODN (Zhu et al., 2019)	3.91	5.43	3.95
DeCaFa (Extra Data) (Dapogny et al., 2019)	2.93	5.26	3.39
HRNet (Wang et al., 2020)	**2.91**	**5.11**	**3.34**
3FabRec (Browatzki & Wallraven, 2020)	3.36	5.74	3.82
AnchorFace (Xu et al., 2021)	3.12	6.19	3.72
LDDMM-Face	3.07	5.40	3.53
LDDMM-Face (Sparse-LF: 50%)	3.18	5.65	3.67

Table 9: NME_l results on six different categories of challenging cases in WFLW, when trained/predicted both on WFLW. "Sparse-LF" is short for training landmark fraction in sparsely-supervised face alignment.

Method	NME_l (%)						
	Test	Pose	Expression	Illumination	Make-up	Occlusion	Blur
CFSS (Zhu et al., 2015)	9.07	21.36	10.09	8.30	8.74	11.76	9.96
DVLN (Wu & Yang, 2017)	6.08	11.54	6.78	5.73	5.98	7.33	6.88
SAN (Dong et al., 2018)	5.22	10.39	5.71	5.19	5.49	6.83	5.80
LAB (Extra Data) (Wu et al., 2018)	5.27	10.24	5.51	5.23	5.15	6.79	6.32
AVS (Qian et al., 2019)	5.25	9.10	5.83	4.93	5.47	6.26	5.86
HRNet (Wang et al., 2020)	**4.60**	**7.90**	**4.82**	4.60	**4.28**	5.45	5.39
AnchorFace (Xu et al., 2021)	4.62	8.10	5.05	**4.52**	4.47	5.38	**5.12**
LDDMM-Face	4.63	8.21	5.00	4.53	4.31	**5.37**	5.22
LDDMM-Face (Sparse-LF: 50%)	4.79	8.56	5.22	4.67	4.44	5.49	5.36

Table 10: NME_l results on the AFLW test set.

Method	NME_l(%)
CFSS (Zhu et al., 2015)	3.92
SAN (Dong et al., 2018)	1.91
LAB (Extra Data) (Wu et al., 2018)	1.85
ODN (Zhu et al., 2019)	1.63
3FabRec (Browatzki & Wallraven, 2020)	1.84
HRNet (Wang et al., 2020)	1.57
AnchorFace (Xu et al., 2021)	**1.56**
LDDMM-Face	1.65

E Qualitative Evaluation

Comprehensive qualitative results are illustrated in this section. LDDMM-Face exhibits outstanding performance in cross-annotation face alignment in a sparse-to-dense manner, as clearly shown in Figure 6, Figure 7, Figure 8 and Figure 9. Results of cross-annotation face alignment across different datasets are presented in Figure 10, Figure 11, Figure 12 and Figure 13.

Figure 6: Qualitative results of sparsely-supervised LDDMM-Face on the 300W full set (trained with 50% landmarks, tested with full landmarks). Green landmarks are the ones involved in the training annotation scheme. Red landmarks are extra ones predicted by LDDMM-Face.

Figure 7: Qualitative results of sparsely-supervised LDDMM-Face on the HELEN test set (trained with 50% landmarks, tested with full landmarks). Green landmarks are the ones involved in the training annotation scheme. Red landmarks are extra ones predicted by LDDMM-Face.

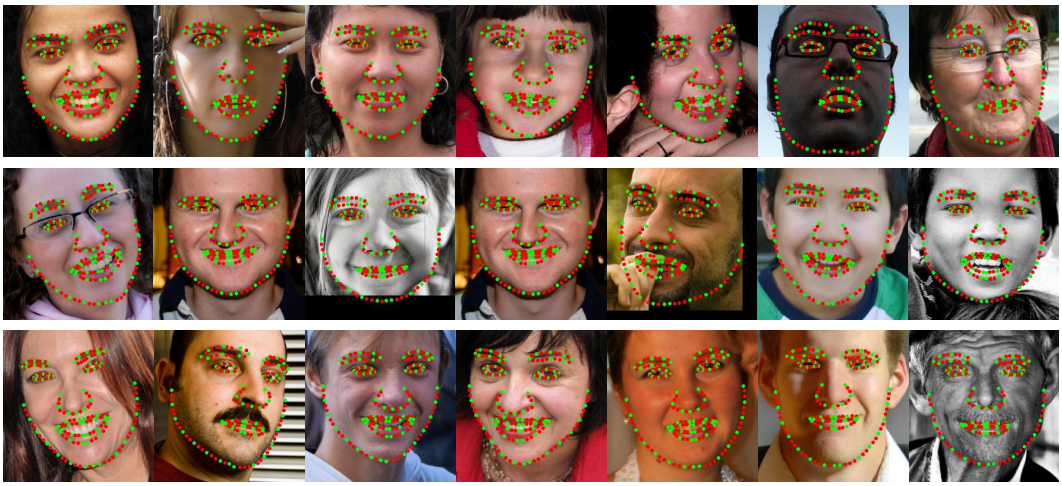

Figure 8: Qualitative results of sparsely-supervised LDDMM-Face on the HELEN test set (trained with 33% landmarks, tested with full landmarks). Green landmarks are the ones involved in the training annotation scheme. Red landmarks are extra ones predicted by LDDMM-Face.

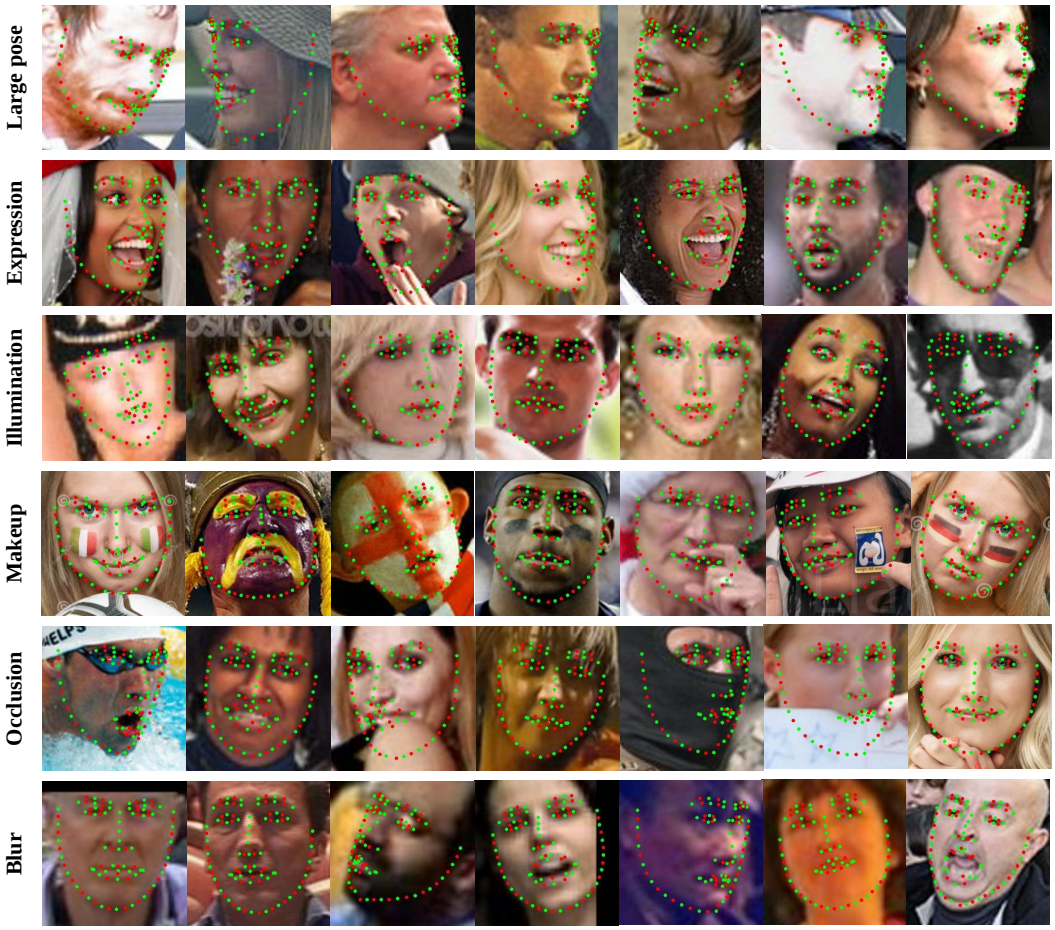

Figure 9: Qualitative results of sparsely-supervised LDDMM-Face on WFLW with six different categories of challenging cases (trained with 50% landmarks, tested with full landmarks). Green landmarks are the ones involved in the training annotation scheme. Red landmarks are extra ones predicted by LDDMM-Face.

Figure 10: Qualitative cross-annotation&cross-dataset comparison results between HRNet and LDDMM-Face on the HELEN test set (trained on 300W, tested on HELEN).

Figure 11: Qualitative cross-annotation&cross-dataset comparison results between HRNet and LDDMM-Face on the 300W full set (trained on HELEN, tested on 300W).

Figure 12: Qualitative cross-annotation&cross-dataset comparison results between HRNet and LDDMM-Face on the WFLW test set (trained on 300W, tested on WFLW).

Figure 13: Qualitative cross-annotation&cross-dataset comparison results between HRNet and LDDMM-Face on the 300W full set (trained on WFLW, tested on 300W).