Multi-Domain Multi-Definition Landmark Localization for Small Datasets

David Ferman[1,2] and Gaurav Bharaj[1]

[1] AI Foundation, USA
[2] UT Austin, USA
davidcferman@gmail.com

Abstract. We present a novel method for multi image domain and multi-landmark definition learning for small dataset facial localization. Training a small dataset alongside a large(r) dataset helps with robust learning for the former, and provides a universal mechanism for facial landmark localization for new and/or smaller standard datasets. To this end, we propose a Vision Transformer encoder with a novel decoder with a definition agnostic shared landmark semantic group structured prior, that is learnt, as we train on more than one dataset concurrently. Due to our novel definition agnostic group prior the datasets may vary in landmark definitions and domains. During the decoder stage we use cross- and self-attention, whose output is later fed into domain/definition specific heads that minimize a Laplacian-log-likelihood loss. We achieve state-of-the-art performance on standard landmark localization datasets such as `COFW` and `WFLW`, when trained with a bigger dataset. We also show state-of-the-art performance on several varied image domain small datasets for animals, caricatures, and facial portrait paintings. Further, we contribute a small dataset (150 images) of pareidolias to show efficacy of our method. Finally, we provide several analysis and ablation studies to justify our claims.

Keywords: Landmarks · Multi-Domain Learning · Vision Transformers

1 Introduction

With the rising need for novel AR/VR, telepresence, character animation filter applications (e.g., adding props and effects in live video streams of humans, pets, etc.), arises the need for facial localization for multiple image domains. While, supervised landmark localization has made great strides for the *in-the-wild* human faces domain, it is often hard to create such datasets for new image domains – animals Khan *et al.* [22], art [49], cartoons, and more recently, pareidolias Song *et al.* [39] that abstractly resemble human faces, Wardle *et al.* [46]. Building a dataset for supervised learning of landmarks is hard due to the cumbersome hand-labeling process, where, hand-labels can lead to noisy and inconsistent landmarks [11], and is often very time consuming for new domains [3].

[3] Labeling a landmark dataset for animal faces can take up to 6,833 hours [22]

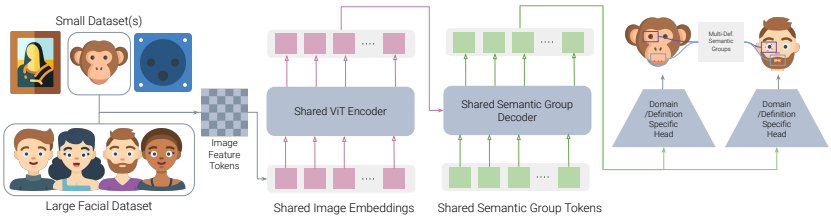

Fig. 1. System overview (left to right): Our method takes a small dataset with a landmark definition, and a larger facial dataset with a different landmark definition, and relies on common semantic group definitions to learn for both dataset concurrently.

Due to varied new domains, Figure 2, and subsequent specific applications, there's no preset definition for facial landmarks. For example, a landmark definition set that works for humans faces may not work for animal faces and vice-versa and thus makes cross-domain learning infeasible. Additionally, within human face localization problems, different datasets have different definitions of landmarks, see Figure 2 (Humans), and certain applications can require unique landmark definitions, e.g., landmarks which correspond to mesh vertices, Wei *et al.* [49]. The landmark datasets necessitated by a particular new application are either small or non-existent. As a result, novel applications that need localization for new image domains and/or definitions becomes infeasible. Image domain localization problems have been previously approached with domain transfer methods. For example, Yaniv *et al.* [56] use domain transfer to approach learning for facial portrait artwork, Wei *et al.* [49] learn landmark correspondences as an auxiliary aspect of mesh fitting. Such methods need a specialized larger dataset and/or landmark definitions from a previous dataset, which might be sub-optimal for the candidate domain. In this work, we create a method that learns robust landmarks for new domains for which small datasets may exist, or for which a small set of labeled images can be obtained, inexpensively, while being landmark definition agnostic.

Poggio *et al.* [33] and White *et al.*[50] observe that shapes share abstract similarities while domains vary. Inspired by this observation and unlike most landmark localization methods [37,48,4,45] our approach models shared *abstract similarities*, i.e., learns together groups of facial landmark semantic groups, Fig 2, rather than learn landmarks directly. The facial landmark semantic group learning can be shared across domains and definitions. Thus, while image domains and localization definitions vary, learning a single representation for each semantic facial group enables generalization of learning across domains and definitions.

Transformers [44] were introduced for natural language processing problems, that model word sequences, e.g., "[The] [quick] [brown] [fox] ...", as densely meaningful tokenized vectors. These vectors are initially indexed from a learned embedding matrix which captures each token's definition, learned across training instances, while instance-specific representations are built contextually via a series of attention layers. Inspired by the success of transformers in NLP, the flexible

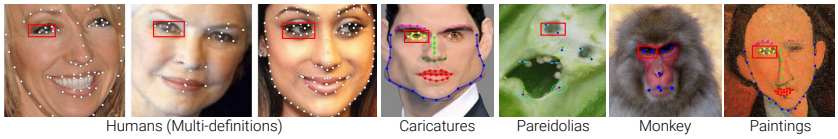

Fig. 2. Our method works for multiple domains and multiple definitions of facial landmark localization problems. The red box represents a *landmark semantic group* shared across different domains and landmark definitions.

handling of multiple tasks and language domains, Raffel *et al.* [35], we consider modeling faces analogously, as a fixed "sentence" of tokens representing facial landmark semantic groups. We seek for our model to learn general definitions via semantic group embeddings, as an implicit facial prior, for predicting semantic group information from image feature contexts.

To this end, we design a novel vision transformer (ViT) [12] architecture for the multiple domain and multi facial landmark definition localization problem. As shown in Figure 1, we first pass the image through our ViT encoder to obtain image feature tokens. These tokens are fed into our novel facial landmark semantic group decoder, which builds contextualized representations of semantic group tokens via cross-attention with image feature tokens and inter-group self-attention. Finally, each facial landmark semantic group vector is passed through definition/dataset specific heads to regress to the final landmarks. Thus, our method treats the facial landmark semantic groups in a general manner, while being capable of predicting landmarks for a variety of domains and definition.

We train our model in a *multi-domain multi-dataset* fashion for small datasets and achieve state-of-the-art performance on `COFW` [5], a small dataset with only 1345 images and a 29 landmark definition and very competitive performance on `WFLW` [52]. Additionally, we display our method's versatility in adapting to very small (roughly 100 image) datasets of animals (monkeys), caricatures, artwork image domains, and contribute a small novel dataset for pareidolias. We show great improvements via multi-domain multi-dataset learning with ablation, qualitative, and quantitative analysis. To summarize, our contributions include:

1. We introduce multi-domain multi-definition learning for the small dataset facial landmark localization problem.
2. We introduce a novel vision transformer encoder-decoder architecture which enables multi-domain (dataset) multi-definition learning via decoding landmark information via shared facial component queries in the decoder.
3. Our method achieves state-of-the-art performance on standard multiple domain facial localization datasets, along with never before seen facial localization domain small datasets, such as, pareidolias.

2 Related Works

Multi-Domain Learning. Multi-domain learning predicts instance labels given both instance features and domain labels, where the goal is to learn a model that improves over a baseline that trains solely on the domain [21,51]. Similar to our work, several studies utilize multi-domain learning to boost performance on a domain with few labeled examples via concurrent training with a domain with plentiful labels [2,51,13,15]. Joshi *et al.* [21] note two approaches for multi-domain learning: domain-specific parameters and modeling inter-domain relations; our approach utilizes both simultaneously. For the image classification task, Dvornik *et al.*[13] propose a feature selection approach, while Zheng *et al.* [59] propose a domain confusion loss to encourage the network to learn domain invariant representations for image classification [15]. For facial landmarks, there may exist domain-specific biases in the outputs between domains, so this property is less desired [56]. Most similar to our approach, Nam *et al.* [30] introduce multi-domain learning for sequence tracking, where their network shares weights for the bulk of the architecture, with domain specific final layers. Our approach utilizes separate final layers for each domain, while exploiting the relations between domains in our decoder by learning shared representations for facial components.

Multi-Definition Learning. The multi-definition problem for facial landmarks solves for inconsistencies between landmark labels to improve model robustness via multi-dataset training [53]. Multi-definition learning is similar to multi-domain in the sense that there is a target dataset for which performance is optimized with shared learning from a source dataset [38]. Smith *et al.* [38] propose to predict a super-set of landmark definitions, while Zhu *et al.* [60] propose an alignment module to estimate pseudo-labels in schema of a target dataset. Motivated by cross-dataset input variation and definition mismatch, Zhang *et al.* [57] propose an intermediate shape regression module that regresses shared sparse definitions that helps inform final regression to the landmark super-set. Wu *et al.* [53] utilize a shared CNN-backbone, prior to dataset/definition specific final direct regression heads. As recent state-of-the-art methods have been heatmap-based, Zhu *et al.* [60] propose separate definition-specific heatmap decoders that tightly couple the decoder architectures with output heatmap definitions [20]. Our method shares abstract similarity to [57]'s shape regression. We include sparse intermediate predictions that are latent vectors rather than explicit landmarks. Similar to [57,53], we also employ definition-specific regression heads, see Section 3.

CNN and Heatmap-based Landmark Learning. Wei *et al.* [48] introduce heatmap-based estimation of 2D landmarks for human pose estimation, later Kowalski *et al.* [24] adapt it for facial landmarks. While heatmaps provide intrinsic spatial generalization [32], they induce quantization errors [18,3,26]. Stacked hourglass networks [31,4,55] or multi-scale processing [40] are then used for building global context. Jin *et al.* [18] note that connecting CNN features to fully connected layers provides a global predictive capacity that leads to inaccurate predictions due to immediate spatial connections, however, this does lead to more consistent predictions.

CoordConv [27] connect CNN features with positional information by injecting a fixed spatial bias through two additional image channels that provide global positional information of {x, y} coordinates respectively. It was adopted by previous state-of-the-art [45] and LAB [52] to capture global information in CNNs via boundary heatmaps that connect semantic groups of landmarks, e.g., eyes, mouth, etc., into semantically grouped heatmaps on a single global boundary heatmap. Chandran *et al.* [8] propose a hard-attention cropping derived from an initial global pass to consider each semantic region of the face and obtain regional heatmaps for each region for high-resolution images.

Transformers for Landmark Learning. Transformers [44] were introduced for vision tasks by DETR [7]'s use of a transformer encoder-decoder over CNN-encoded features for the object detection. Vision Transformers (ViT) [12] show promising performance for vision tasks without the use of CNNs, while DEiT [42] use a CNN for knowledge distillation for further improvements. Swin [29], inspired by CNN architectures, propose a hierarchically processed shifted window attention approach. However, we adopt a simple vanilla ViT [12], and employ a transformer decoder for predicting the latent landmark information for facial semantic group regression.

HiH [26] resolve for heatmap quantization errors and study a CNN-based versus transformer-based heatmap prediction network with a CNN-backbone. LOTR [47] show that transformers can be used to break the direct spatial dependencies induced by CNN-MLP architectures for performant direct regression. They employ a CNN-backbone followed by a transformer-encoder decoder, where the decoder queries correspond to individual landmarks, followed by MLP regression heads. Recently, FarRL [1] introduce a BERT/BEiT-like transformer pre-training equivalent on faces, pre-training self-supervisedly on 20 million facial images, and predicting facial landmarks, with heatmap prediction, as one of three facial tasks in a multi-task setup.

Our method combines transformer-based (cross and self-attention) direct regression with the semantic group intuition of LAB, as our novel transformer decoder predicts representations for the semantic groups prior to explicitly regressing landmarks contained in the semantic group. Our method is most similar to LOTR, except that while LOTR is DETR-like [7] with its full CNN-backbone, our method is ViT-like, using projection patchification, and no CNN feature backbone. Also, LOTR queries each individual landmark from the encoded image features, whereas our method queries semantic landmark groupings, e.g., nose, for multi-domain/definition learning purposes, and regresses both landmark mean and covariance information.

Multi-Dataset Learning. In order to address the small-datasets common among facial landmark problems, several approaches have been devised. These include semi-supervised learning [16,3], self-supervised learning [59], and multi-dataset learning [3,59]. Zheng *et al.* [59], inspired by BERT-inspired [10] BEiT [1], use self-supervised pre-training techniques to learn general facial representations, employing both a contrastive learning approach using textual labels as well as a masked image prediction methodology on 20 million facial images. Qian*et al.* [34]

introduce a synthetic data creation methodology which employs self-supervised learning to translate labeled faces into the style of other images, achieving large performance boosts over purely supervised methods. Jin *et al.* [19] introduce cross-protocol network training, where multiple facial landmark datasets are trained simultaneously by sharing a backbone feature encoder and using a different heatmap decoder network for each dataset and thus only shares weights in the CNN feature backbone, but not in the landmark heatmap decoders. Our work is similar to Jin *et al.* in that we train on multi-definition facial landmark datasets as our model's source of additional data. However, rather than decoding each dataset separately, our definition agnostic decoder shares weights across datasets by modeling shared semantic groupings of landmarks.

3 Method

Our goal is to create a robust landmark localization solution for small data regime problems, where due to varied circumstances acquiring a large dataset is infeasible or expensive. We approach this problem via a multi-domain multi-definition (MDMD) facial landmark localization formulation with a transformer-based encoder-decoder architecture. The MDMD problem consists of predicting facial landmarks for target image domain(s) or landmark definition(s), while training on one or more source domains/definitions. Our novel transformer-based architecture is shown in Fig 1. We select n image domain datasets, which may have different landmark definitions as our MDMD input. The various landmark definitions map to a standard semantic grouping which we define for each dataset, Sec. 3.2. These n domain datasets are then fed into a ViT [12] encoder that builds image feature representations from the input images, Sec. 3.1. Predefined shared semantic group tokens act as a learnt structure "prior" to the shared semantic group decoder, that takes as input the encoder's output feature tokens. Then, the decoder builds representations of these definition-agnostic semantic groups by attending to both the image features via cross-attention and the other groups via self-attention, Sec. 3.3. Finally, the semantic group tokens – output of the decoder, each of which individually correspond to a unique set of landmarks (out of n), are then fed into regression heads that predict the final landmarks, Sec 3.4.

3.1 ViT Encoder

We employ a pre-trained ViT [12] encoder that is shared across images from all input domains. The model first patchifies the input image to transform the initial image $I \in \mathbb{R}^{224\times224\times3}$ into a grid $G \in \mathbb{R}^{14\times14\times D}$, with $D = 768$, and is then flattened, appended with a global token, and combined with positional encodings to obtain the ViT input tokens $X_{in} \in \mathbb{R}^{(196+1)\times D}$. The input tokens X are then passed through a series of ViT layers consisting of self-attention and MLPs. The final feature tokens are then obtained and passed to the shared decoder to extract landmark information from these generic facial features.

3.2 Facial Landmark Semantic Group (FLSG)

In-order to have a universal mechanism for supporting various landmark definitions, we propose a novel facial landmarks semantic group prior. This abstract view of the face leads to definition and domain generalization through the relaxation of strict spatial dependencies, such as in boundary heatmaps [52]. We divide the facial landmarks for each definition into a set of 12 shared FLSGs, as shown in Figure 3. FLSGs are modeled as an embedding matrix $F_{in} \in \mathbb{R}^{12\times D}$, where each $F_{in}^{(i)} \in \mathbb{R}^{D}$ represents learned prior information for a particular semantic group of facial landmarks. The decoder exploits the learned FLSG representations F_{in} that are used to initialize the FLSG tokens that act as input to the decoder. While, different FLSGs may have a different number landmarks depending on the definitions, our model does not explicitly differentiate between them until the final prediction head stage, Section 3.4. Thus, our FLSGs unlock MDMD learning via a standard general facial representation.

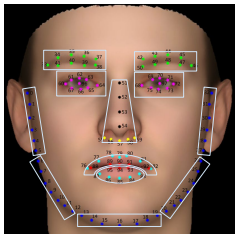

Fig. 3. Facial Landmark Semantic Group. Image source: [52]

3.3 Definition Agnostic Decoder

We want FLSGs to collect information from the image features that it deems relevant (cross-attention) and also collect information about its context wrt other FLSGs (self-attention). We achieve this via a novel definition agnostic decoder, where given initial FLSG tokens $F_{in} \in \mathbb{R}^{12\times D}$ and the encoded image feature tokens $X_{out} \in \mathbb{R}^{(196+1)\times D}$, the decoder seeks to infuse the "structured prior" FLSG tokens with information from both the input image and other FLSG tokens. The decoder is composed of three decoder blocks that consist of self- and cross-attention. Each decoder block contains cross-attention in which the FLSG tokens act as "queries" and the image features as "keys" and "values" [44]. This is followed by self-attention, where the FLSG tokens can perform message passing. Explicitly, given the input FLSG tokens, F_{in}, each decoder block is as follows:

$$F^1_{hidden} = \mathsf{MHCA}(\mathsf{LN}(F_{in}), \mathsf{LN}(X_{out}))) + F_{in} \tag{1}$$

$$F^2_{hidden} = \mathsf{MHSA}(\mathsf{LN}(F^1_{hidden})) + F^1_{hidden} \tag{2}$$

$$F_{out} = \mathsf{FFN}(\mathsf{LN}(F^2_{hidden})) \tag{3}$$

where MHCA, MHSA, LN, and FFN are the standard transformer multi-head cross-attention, multi-head self-attention, layer normalization, and feed-forward network respectively [44]. Thus, decoder layers infuse the FLSG tokens with image feature information as well as inter-FLSG contextual information, so that they contain information pertaining to localizing the landmarks contained in the given semantic group. The final FLSG tokens $F_{out} \in \mathbb{R}^{12\times D}$, output by decoder, are then plugged into the definition specific prediction heads.

3.4 Definition/Domain-Specific Prediction Heads

Finally, we employ definition/domain specific prediction heads, that directly regress the landmarks that correspond to each FLSG. An image I_j is provided to our model, where j is the dataset (definition) index. The dataset index is simply used to route the FLSG tokens to the head that corresponds to that dataset (see pseudocode in the supplementary). Each dataset's landmark head, regresses from an FLSG vector $F^i_{out} \in \mathbb{R}^D$ via two-layer $\mathsf{MLP}_{\mathsf{lm}}{}^i_j$ to output landmarks $L^i_j \in \mathbb{R}^{N^i_j \times 2}$, where N^i_j is the number of landmarks for the ith FLSG and the jth dataset.

Rather than predict landmarks alone, following Kumar *et al.* [25], we also predict the covariance information via a Cholesky estimation head, $\mathsf{MLP}_{\mathsf{chol}}{}^i_j$, obtaining a second output corresponding to each FLSG, $C^i_j \in \mathbb{R}^{N^i_j \times 3}$, corresponding to the parameters of the Cholesky factorization of a predicted covariance matrix. While Kumar *et al.*'s Cholesky estimation network regresses from the latent bottleneck vector of their CNN, and use heatmap-based prediction for the mean estimate, supervising outputs from several stacked hourglass layers of their DU-NET [41], we utilize a shared FLSG vector to predict both mean and covariance information. The final minimization loss function we use to train our model is as follows:

$$\mathcal{L}_{\mathsf{MDMD}} = \frac{1}{|\mathsf{FLSG}|} \sum_{i=1}^{|\mathsf{FLSG}|} \left[\frac{1}{N^i_j} \sum_{k=1}^{N^i_j} \mathcal{L}_{lll}(L^i_j, C^i_j, {L_{GT}}^i_j)_k \right] \tag{4}$$

Here, L_{lll} is the Laplacian log-likelihood (see supplementary) and ${L_{GT}}^i_j$ is the ground truth landmarks.

4 Experiments and Results

We evaluate our model's multi-domain and multi-definition learning capabilities on novel domains with small datasets: `AnimWeb` [22], `ArtFace` [56], `CariFace` [58] and `PARE` dataset [New], as well as standard benchmark datasets: `COFW` [5] and `WFLW` [52] (and `300W` [36], `LaPa` [28]), see supplementary material for details on datasets.

For each experiment, we report normalized mean error (NME) with inter-ocular normalization as well as inter-pupil, where comparison necessitates. Additionally, we report Area Under The Curve (AUC) and FR (Failure Rate) scores, considering a failure as mean NME greater than 10% for a given face. While 256×256 input crops are most commonly used [34], our ViT [12] encoder was pre-trained with 224×224 input crops, that we adopt. All models are trained with the Adam[23] optimizer with learning rate $1e^{-4}$ and linear learning rate decay. For each experiment, we consider performance for our model training with a single domain and definition, and then compare its performance when training with an additional dataset in the multi-domain and multi-definition fashion. In order to train concurrently across datasets, for each mini-batch, we uniformly sample a

Fig. 4. Qualitative results for our method across several datasets. Key: GT landmarks, predicted landmarks, error vectors, uncertainty estimation

dataset from which we draw batch samples. We include additional implementation details, including augmentation strategy, in the supplementary materials. In the following, we discuss various qualitative (Figure 4), and quantitative results on various datasets:

4.1 COFW [5]

We evaluate our method on the COFW dataset that contains 1,345 training images, and 500 testing images. We note that among standard benchmark datasets, COFW is most similar to our problem for its unique 29 landmark definition as well as its relatively small size. We train our model with two settings: COFW, and COFW concurrently trained with LaPa. We evaluate our model with inter-pupil normalization, following [45,17], surpassing state-of-the-art, Table 1. We also note that for each dataset on which we train our model, concurrent training with a larger dataset shows significant performance improvements.

Table 1. Comparison against SOTA for COFW [5]

Method	$NME_{ip}(\%)$	$FR_{10\%}$	$AUC_{10\%}$
Wing [14]	5.44	3.75	-
DCFE [43]	5.27	7.29	35.86
AWing [45]	4.94	.99	48.82
ADNet [17]	4.68	.59	53.17
MDMD Base	4.82	**.39**	51.84
MDMD w/LaPa	**4.65**	.59	**53.49**

4.2 WFLW [52]

We further evaluate our method on the WFLW dataset, which consists of 7,500 training images and 2,500 testing images, with a 98 landmark definition. We train our model in the multi-definition manner with two settings: WFLW and WFLW concurrently with LaPa, where LaPa presents 19,000 faces with a 106 landmark definition. As 300W and COFW are relatively small with 3837 and 1345 training faces respectively, we do not consider the concurrently training with these smaller datasets, as this runs contrary to our goal of boosting performance from training with larger datasets. We compare our results with other methods for NME, FR, and AUC on the full test set along with subsets which test for robustness on large poses, expression, illumination, make-up, occlusion, and blur, in Table 2. Our method outperforms all previous state-of-the-art methods for overall scores aside from two concurrent works [3,59]. Our method also achieves SOTA performance compared to previously reported methods for the majority of subsets for NME, FR, and AUC. See qualitative comparisons for our method in Figure 4.

4.3 Small Dataset Experiments

We consider our methods performance for small datasets of novel domains and landmark definitions. For each of these experiments, we train both a baseline model on the small dataset only as well as a multi-domain and multi-definition model, for which we employ the moderately sized 300W dataset. Per Williams *et al.* [51], the goal of multi-domain learning is to show improvement over a single domain baseline. While for previous experiments, our focus was primarily on how our method compares with previous methods, here, we compare against a baseline single-domain training. Where applicable we draw rough comparisons with previous works for these datasets. We report relative NME, FR, and AUC for all small dataset experiments in Table 3. We observe large performance gains for each dataset through the generalized learning via our multi-domain and multi-definition approach.

AnimWeb [22]

While the AnimWeb [22] dataset features 21,900 animal faces across 334 species, we select a single specie, the Japanese Macaque, containing 133 examples which we split into 100 training and 33 testing monkey faces for our experiment. We

Table 2. Comparison against SOTA for `WFLW` [52]. *Concurrent works, Key: best, second

Metric	Method	Testset	Pose Subset	Expression Subset	Illumination Subset	Make-up Subset	Occlusion Subset	Blur Subset
NME(%)	ESR [6]	11.13	25.88	11.47	10.49	11.05	13.75	12.20
	SDM [54]	10.29	24.10	11.45	9.32	9.38	13.03	11.28
	CFSS [61]	9.07	21.36	10.09	8.30	8.74	11.76	9.96
	DVLN [53]	6.08	11.54	6.78	5.73	5.98	7.33	6.88
	LAB [52]	5.27	10.24	5.51	5.23	5.15	6.79	6.12
	Wing [14]	5.11	8.75	5.36	4.93	5.41	6.37	5.81
	DeCaFA [9]	4.62	8.11	4.65	4.41	4.63	5.74	5.38
	AWing [45]	4.36	7.38	4.58	4.32	4.27	5.19	4.96
	LUVLi [25]	4.37	-	-	-	-	-	-
	AWing [45]	4.36	7.38	4.58	4.32	4.27	5.19	4.96
	HiH [26]	4.18	7.20	4.19	4.45	3.97	5.00	4.81
	ADNet [17]	4.14	6.96	4.38	4.09	4.05	5.06	4.79
	FaRL [59]*	3.96	6.91	4.21	3.97	3.80	4.71	4.57
	SH-FAN Base [3]*	4.20	-	-	-	-	-	
	SH-FAN [3]*	3.72	-	-	-	-	-	
	MDMD Base*	4.06	7.11	4.21	3.88	4.04	4.86	4.63
	MDMD w/LaPa*	3.97	6.90	4.11	3.80	3.90	4.78	4.49
FR_{10}(%)	ESR [6]	35.24	90.18	42.04	30.80	38.84	47.28	41.40
	SDM [54]	29.40	84.36	33.44	26.22	27.67	41.85	35.32
	CFSS [61]	20.56	66.26	23.25	17.34	21.84	32.88	23.67
	DVLN [53]	10.84	46.93	11.15	7.31	11.65	16.30	13.71
	LAB [52]	7.56	28.83	6.37	6.73	7.77	13.72	10.74
	Wing [14]	6.00	22.70	4.78	4.30	7.77	12.50	7.76
	DeCaFA [9]	4.84	21.40	3.73	3.22	6.15	9.26	6.61
	AWing [45]	2.84	13.50	2.23	2.58	2.91	5.98	3.75
	LUVLi [25]	3.12	-	-	-	-	-	-
	HiH [26]	2.84	14.41	2.55	2.15	1.46	5.71	3.49
	ADNet [17]	2.72	12.72	2.15	2.44	1.94	5.79	3.54
	FaRL [59]*	1.76	-	-	-	-	-	-
	SH-FAN [3]*	1.55	-	-	-	-	-	-
	MDMD Base*	2.63	14.11	1.91	1.71	2.43	4.89	2.98
	MDMD w/LaPa*	2.2	11.96	1.27	1.58	1.46	4.35	2.59
$AUC_{10\%}$	ESR [6]	0.2774	0.0177	0.1981	0.2953	0.2485	0.1946	0.2204
	SDM [54]	0.3002	0.0226	0.2293	0.3237	0.3125	0.2060	0.2398
	CFSS [61]	0.3659	0.0632	0.3157	0.3854	0.3691	0.2688	0.3037
	DVLN [53]	0.4551	0.1474	0.3889	0.4743	0.4494	0.3794	0.3973
	LAB [52]	0.5323	0.2345	0.4951	0.5433	0.5394	0.4490	0.4630
	Wing [14]	0.5504	0.3100	0.4959	0.5408	0.5582	0.4885	0.4918
	DeCaFA [9]	0.5630	0.2920	0.5460	0.5790	0.5750	0.4850	0.4940
	AWing [45]	0.5719	0.3120	0.5149	0.5777	0.5715	0.5022	0.5120
	LUVLi [25]	0.5770	-	-	-	-	-	-
	HiH [26]	0.597	0.342	0.590	0.606	0.604	0.527	0.549
	ADNet [17]	0.6022	0.3441	0.5234	0.5805	0.6007	0.5295	0.5480
	FaRL [59]*	0.6116	-	-	-	-	-	-
	SH-FAN [3]*	.6310	-	-	-	-	-	-
	MDMD Base*	.6010	.3316	.5870	.6179	.5998	.4978	.5476
	MDMD w/LaPa*	.6083	.3438	.5933	.6252	.6127	.5354	.5582

Table 3. Evaluation of multi-domain and multi-definition learning capabilities across small datasets for novel domains and landmark definitions

Dataset	Method	NME_{ic}	$FR_{10\%}$	$AUC_{10\%}$	# Landmarks
`AnimWeb` [22]	MDMD Base	6.88	**15.15**	.4233	9
	MDMD w/300W	**6.55**	**15.15**	**.4388**	9
`ArtFace` [56]	MDMD Base	4.46	2.08	.5549	68
	MDMD w/300W	**3.75**	**0.0**	**.63**	68
`CariFace` [58]	MDMD Base	7.81	19.04	.2941	68
	MDMD w/300W	**5.85**	**6.41**	**.4357**	68
`PARE`	MDMD Base	9.12	28.0	.2365	9
	MDMD w/300W	**8.59**	**22.0**	**.2871**	9

Table 4. Comparison against previous work for `AnimWeb` [22] (left) and `ArtFace` [56] (middle). Following Khan, NME scores for `AnimWeb` are normalized by bounding box size. Comparison against previous work for `CariFace` [58](right).

Method	NME_{box}
Khan *et al.* [22]	5.23
MDMD Base	3.66
MDMD 300W	**3.44**

Method	NME_{ic}
Yaniv *et al.* [56]	4.52[2]
MDMD Base	4.46
MDMD 300W	**3.72**

Method	NME	Trn Imgs
Zhang*et al.* [58]	**5.83**	6,420
MDMD Base	7.81	148
MDMD 300W	5.85	148

train jointly between `300W` and the monkey domains and definitions. We note that the animals are labeled with 9 landmarks, while `300W` is labeled with 68. For comparison against previous work, we cannot compare directly, as Khan *et al.* [22] train on a variety of species on a dataset with significantly large amount of training data. Furthermore, their scores represent a wide variety of animals, while ours are a subset of just one animal. Nevertheless, we report our scores for ballpark comparison in Table 4 (left).

`ArtFace` [56]

We compare our model's performance on 'faces in artworks' domain, utilizing the `ArtFace` [56] dataset, consisting of 160 faces, 10 faces per 16 artists, with a `300W`-like 68 landmark definition. While Yaniv *et al.* [56] utilize an elaborate geometric-aware and style transfer to augment `300W` images for training, and perform generalization to the artworks domain for testing, we split each artist by taking the first 7 image indices for training with the other 3 for testing. [4]

`CariFace` [58]

Zhang *et al.* [58] introduce an interesting problem of localizing landmarks on the domain of human caricatures, introducing the `CariFace` dataset that they train on 6,240 images and test on 1,560. Rather than train on the full set, we simulate the small dataset problem for this novel domain by taking the first 148 images for training and evaluating on the full test set. Similar to `ArtFace` [56], `CariFace` [58] uses the same `300W` landmark definition. We compare our NME

[4] We compare our results against previous work, with a caveat that our evaluation is on a subset of the dataset rather than the full dataset, and achieve SOTA performance for the `ArtFace` dataset, as shown in Table 4 (right).

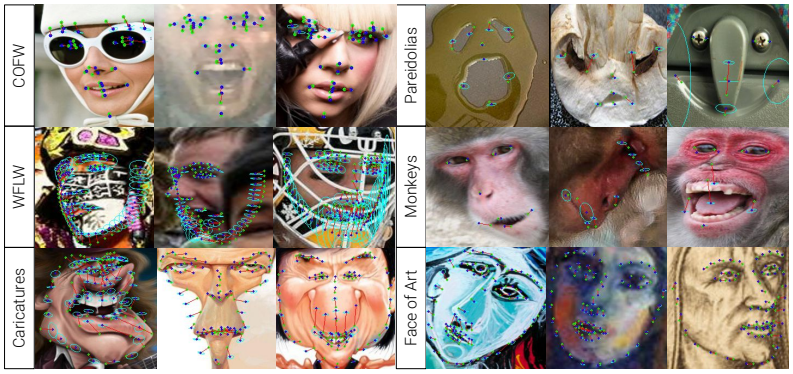

Fig. 5. Qualitative results for displaying the limitations our method across several datasets. Key: GT landmarks, predicted landmarks, error vectors, uncertainty estimation

Table 5. Ablation studies.

Method	NME_{ip}	$FR_{10\%}$	$AUC_{10\%}$
MDMD Single w/Euclidean loss	4.90	.79	.5100
MDMD Single w/landmark tokens	4.73	.59	.5278
MDMD Single	4.82	.59	.5184
MDMD w/LaPa	**4.64**	**.39**	**.5349**

scores trained on 40X less data from the caricature domain, and achieve slightly lower performance than Zhang *et al.*, as shown in Table 4.

`PARE`

Finally, we consider a unique dataset of illusory faces, also known as pareidolias, that we obtained from [46], and labeled 150 images with 9 landmarks each. This domain is particularly interesting, as the face pictures are only abstractly similar to the human faces from the `300W` dataset with which the model trains concurrently. As shown in Table 3, performance greatly improves with multi-domain and multi-definition learning.

4.4 Ablation Analysis

In addition to training with and without an additional dataset, we perform ablation studies for a several architectural components of our model. For each study, we test performance on the `COFW` dataset alone. First, we remove our facial landmark semantic grouping tokens from our decoder, and replace them with individual landmark tokens. Next, we train with simple Euclidean loss, rather than Lapalacian log-likelihood. We show our comparisons for against the baseline model in Table 5.

Small Datasets with v. without `300W` *[36].* We compare our model's performance with and without an additional dataset when training on small datasets of novel domains and definitions. We observe that for each dataset, training without

the additional data leads to severe performance reductions, Table 3. Thus, we conclude that our multi-domain and multi-domain learning strategy is effective at exploiting additional labeled data for small datasets of novel domains and definitions.

Laplacian Log-Likelihood v. Euclidean Loss. To evaluate the effectiveness of our Laplacian log-likelihood training objective, we compare against a simple baseline of Euclidean distance loss. We train our model on COFW [5] and show that performance severely deteriorates when we use Euclidean loss, Table 5.

Facial Landmark Semantic Group (FLSG) v. Explicit Landmark Modeling. Lastly, we seek to evaluate the effectiveness of our FLSG modeling when compared to a simple baseline of modeling each landmark with its own token. As our MDMD method relies on FLSGs to accomplish multi-definition learning, and thus, cannot be removed while still accomplishing the same task, we instead consider its effectiveness when training with a single dataset, COFW [5]. We observe a decrease in performance when training with the FLSG in the standard scenario of a single dataset, Table 5. However, this decrease is overcome by multi-dataset learning. Thus, FLSG acts as a strategy for achieving performance gains in the multi-domain/definition scenario, while landmark queries was better for the single dataset case, in this case.

5 Limitations and Conclusion

We introduced a method for multi-domain and multi-definition landmark localization, that employs a transformer that models facial landmark semantic groups (FLSGs) as opposed to individual landmarks, in-order to share learning across domains and definitions. Our method achieves state-of-the-art performance, as well as successfully improves over baselines of single-domain learning for both small and large datasets. We note however, that our model still struggles with certain difficult circumstances, such as extreme pose and occlusions, as shown in Figure 5. Another limitation is extremely deformed face shapes, for example, the middle caricature face, Fig. 5.

We also note that FLSGs aid in multi-definition learning, but hurt performance for the single dataset scenario (Table 5). Thus, in the future, we want to explore exploitation of explicit landmark modeling jointly with FLSGs to obtain the best of both worlds. We also note that the proposed FLS-Grouping (after several test permutations) works well for all domains/definitions and helps with generalization, in the future we want to explore domain/definition specific grouping. Additional future work may consider extending these ideas to multi-task learning, temporal modeling, or toward zero-shot and few-shot learning.

References

1. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. ArXiv **abs/2106.08254** (2021) 5

2. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F.C., Vaughan, J.W.: A theory of learning from different domains. Machine Learning **79**, 151–175 (2009) 4
3. Bulat, A., Sanchez, E., Tzimiropoulos, G.: Subpixel heatmap regression for facial landmark localization. In: Proceedings of the British Machine Vision Conference (BMVC) (2021) 4, 5, 10, 11
4. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1021–1030 (2017) 2, 4
5. Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: Proceedings of the IEEE international conference on computer vision. pp. 1513–1520 (2013) 3, 8, 9, 10, 14
6. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. International Journal of Computer Vision **107**, 177–190 (2012) 11
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229. Springer (2020) 5
8. Chandran, P., Bradley, D., Gross, M.H., Beeler, T.: Attention-driven cropping for very high resolution facial landmark detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5860–5869 (2020) 5
9. Dapogny, A., Bailly, K., Cord, M.: Decafa: Deep convolutional cascade for face alignment in the wild. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 6892–6900 (2019) 11
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.N.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018), https://arxiv.org/abs/1810.04805 5
11. Dong, X., Yu, S.I., Weng, X., Wei, S.E., Yang, Y., Sheikh, Y.: Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 360–368 (2018) 1
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 3, 5, 6, 8
13. Dvornik, N., Schmid, C., Mairal, J.: Selecting relevant features from a multi-domain representation for few-shot classification. In: ECCV (2020) 4
14. Feng, Z.H., Kittler, J., Awais, M., Huber, P., Wu, X.J.: Wing loss for robust facial landmark localisation with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2235–2245 (2018) 10, 11
15. Hoffman, J., Tzeng, E., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. 2015 IEEE International Conference on Computer Vision (ICCV) pp. 4068–4076 (2015) 4
16. Honari, S., Molchanov, P., Tyree, S., Vincent, P., Pal, C., Kautz, J.: Improving landmark localization with semi-supervised learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1546–1555 (2018) 5
17. Huang, Y., Yang, H., Li, C., Kim, J., Wei, F.: Adnet: Leveraging error-bias towards normal direction in face alignment. arXiv preprint arXiv:2109.05721 (2021) 9, 10, 11

18. Jin, H., Liao, S., Shao, L.: Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. International Journal of Computer Vision **129**(12), 3174–3194 (2021) 4
19. Jin, S., Feng, Z., Yang, W., Kittler, J.: Separable batch normalization for robust facial landmark localization with cross-protocol network training. arXiv preprint arXiv:2101.06663 (2021) 6
20. Jin, S., Feng, Z., Yang, W., Kittler, J.: Separable batch normalization for robust facial landmark localization with cross-protocol network training. ArXiv **abs/2101.06663** (2021) 4
21. Joshi, M., Dredze, M., Cohen, W.W., Rosé, C.P.: Multi-domain learning: When do domains matter? In: EMNLP (2012) 4
22. Khan, M.H., McDonagh, J., Khan, S.H., Shahabuddin, M., Arora, A., Khan, F.S., Shao, L., Tzimiropoulos, G.: Animalweb: A large-scale hierarchical dataset of annotated animal faces. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 6937–6946 (2020) 1, 8, 10, 12
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 8
24. Kowalski, M., Naruniec, J., Trzciński, T.: Deep alignment network: A convolutional neural network for robust face alignment. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 2034–2043 (2017) 4
25. Kumar, A., Marks, T.K., Mou, W., Wang, Y., Jones, M., Cherian, A., Koike-Akino, T., Liu, X., Feng, C.: 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) 8, 11
26. Lan, X., Hu, Q., Cheng, J.: Hih: Towards more accurate face alignment via heatmap in heatmap. arXiv preprint arXiv:2104.03100 (2021) 4, 5, 11
27. Liu, R., Lehman, J., Molino, P., Such, F.P., Frank, E., Sergeev, A., Yosinski, J.: An intriguing failing of convolutional neural networks and the coordconv solution. In: NeurIPS (2018) 5
28. Liu, Y., Shi, H., Si, Y., Shen, H., Wang, X., Mei, T.: A high-efficiency framework for constructing large-scale face parsing benchmark. arXiv preprint arXiv:1905.04830 (2019) 8
29. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021) 5
30. Nam, H., Han, B.: Learning multi-domain convolutional neural networks for visual tracking. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4293–4302 (2016) 4
31. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV (2016) 4
32. Nibali, A., He, Z., Morgan, S., Prendergast, L.: Numerical coordinate regression with convolutional neural networks. ArXiv **abs/1801.07372** (2018) 4
33. Poggio, T., Torre, V., Koch, C.: Computational vision and regularization theory. Readings in computer vision pp. 638–643 (1987) 2
34. Qian, S., Sun, K., Wu, W., Qian, C., Jia, J.: 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) 5, 8

35. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683 (2019) 3
36. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 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 (2013) 8, 13
37. Saragih, J.M., Lucey, S., Cohn, J.F.: Face alignment through subspace constrained mean-shifts. In: 2009 IEEE 12th International Conference on Computer Vision. pp. 1034–1041. Ieee (2009) 2
38. Smith, B.M., Zhang, L.: Collaborative facial landmark localization for transferring annotations across datasets. In: ECCV (2014) 4
39. Song, L., Wu, W., Fu, C., Qian, C., Loy, C.C., He, R.: Everything's talkin': Pareidolia face reenactment. arXiv preprint arXiv:2104.03061 (2021) 1
40. Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., Wang, J.: High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514 (2019) 4
41. Tang, Z., Peng, X., Li, K., Metaxas, D.N.: Towards efficient u-nets: A coupled and quantized approach. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**, 2038–2050 (2020) 8
42. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021) 5
43. Valle, R., Buenaposada, J.M., Valdés, A., Baumela, L.: A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment. In: ECCV (2018) 10
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017) 2, 5, 7
45. Wang, X., Bo, L., Fuxin, L.: Adaptive wing loss for robust face alignment via heatmap regression. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6971–6981 (2019) 2, 5, 9, 10, 11
46. Wardle, S.G., Paranjape, S., Taubert, J., Baker, C.I.: Illusory faces are more likely to be perceived as male than female. Proceedings of the National Academy of Sciences **119**(5) (2022) 1, 13
47. Watchareeruetai, U., Sommanna, B., Jain, S., Noinongyao, P., Ganguly, A., Samacoits, A., Earp, S.W., Sritrakool, N.: Lotr: Face landmark localization using localization transformer. arXiv preprint arXiv:2109.10057 (2021) 5
48. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 4724–4732 (2016) 2, 4
49. Wei, S.E., Saragih, J.M., Simon, T., Harley, A.W., Lombardi, S., Perdoch, M., Hypes, A., Wang, D., Badino, H., Sheikh, Y.: Vr facial animation via multiview image translation. ACM Transactions on Graphics (TOG) **38**, 1 – 16 (2019) 1, 2
50. White, T.: Shared visual abstractions. ArXiv **abs/1912.04217** (2019) 2
51. Williams, J.: Multi-domain learning and generalization in dialog state tracking. In: SIGDIAL Conference (2013) 4, 10
52. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: A boundary-aware face alignment algorithm. In: CVPR (2018) 3, 5, 7, 8, 10, 11
53. Wu, W., Yang, S.: Leveraging intra and inter-dataset variations for robust face alignment. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 2096–2105 (2017) 4, 11

54. Xiong, X., la Torre, F.D.: Supervised descent method and its applications to face alignment. 2013 IEEE Conference on Computer Vision and Pattern Recognition pp. 532–539 (2013) 11
55. Yang, J., Liu, Q., Zhang, K.: Stacked hourglass network for robust facial landmark localisation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 79–87 (2017) 4
56. Yaniv, J., Newman, Y.: The face of art: Landmark detection and geometric style in portraits (2019) 2, 4, 8, 12
57. Zhang, J., Kan, M., Shan, S., Chen, X.: Leveraging datasets with varying annotations for face alignment via deep regression network. 2015 IEEE International Conference on Computer Vision (ICCV) pp. 3801–3809 (2015) 4
58. Zhang, J., Cai, H., Guo, Y., Peng, Z.: Landmark detection and 3d face reconstruction for caricature using a nonlinear parametric model. Graph. Model. **115**, 101103 (2021) 8, 12
59. Zheng, Y., Yang, H., Zhang, T., Bao, J., Chen, D., Huang, Y., Yuan, L., Chen, D., Zeng, M., Wen, F.: General facial representation learning in a visual-linguistic manner. CoRR (2021) 4, 5, 10, 11
60. Zhu, S., Li, C., Loy, C.C., Tang, X.: Transferring landmark annotations for cross-dataset face alignment. ArXiv **abs/1409.0602** (2014) 4
61. Zhu, S., Li, C., Loy, C.C., Tang, X.: Face alignment by coarse-to-fine shape searching. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4998–5006 (2015) 11