Few-shot Geometry-Aware Keypoint Localization

Xingzhe He[1]* Gaurav Bharaj[2] David Ferman[2] Helge Rhodin[1] Pablo Garrido[2]

[1] University of British Columbia [2] Flawless AI

Abstract

Supervised keypoint localization methods rely on large manually labeled image datasets, where objects can deform, articulate, or occlude. However, creating such large keypoint labels is time-consuming and costly, and is often error-prone due to inconsistent labeling. Thus, we desire an approach that can learn keypoint localization with fewer yet consistently annotated images. To this end, we present a novel formulation that learns to localize semantically consistent keypoint definitions, even for occluded regions, for varying object categories. We use a few user-labeled 2D images as input examples, which are extended via self-supervision using a larger unlabeled dataset. Unlike unsupervised methods, the few-shot images act as semantic shape constraints for object localization. Furthermore, we introduce 3D geometry-aware constraints to uplift keypoints, achieving more accurate 2D localization. Our general-purpose formulation paves the way for semantically conditioned generative modeling and attains competitive or state-of-the-art accuracy on several datasets, including human faces, eyes, animals, cars, and never-before-seen mouth interior (teeth) localization tasks, not attempted by the previous few-shot methods.
Project page: https://xingzhehe.github.io/FewShot3DKP/

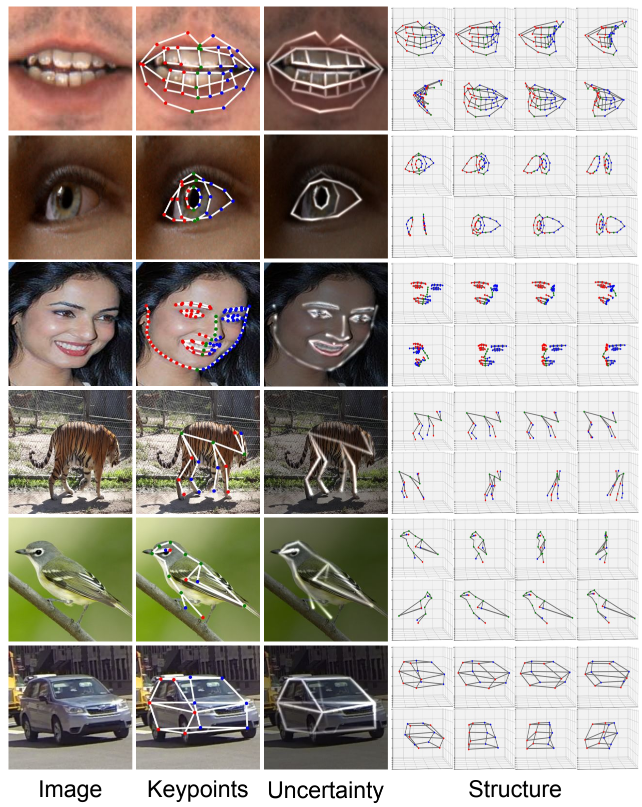

Figure 1. All results are obtained by 10-shot learning except Tigers where 20 examples are used. The left/right/middle keypoints are marked in red/blue/green. Using only a few shots, the model learns semantically consistent and human interpretable keypoints. Uncertainty modeling helps us identify occlusions and ambiguous boundaries, as shown in the mouth, eye, and car example.

1. Introduction

Keypoint localization is a long-standing problem in computer vision with applications in classification [7, 8], image generation [45, 66], character animation [64, 65], 3D modeling [15, 55], and anti-spoofing [9], among others. Traditional supervised keypoint localization approaches require a large dataset of annotated images with balanced data distributions to train robust models that generalize to unseen observations [18, 84, 87]. However, annotating keypoints in images and videos is expensive, and usually requires several annotators with domain expertise [44, 80, 83]. Manual annotations can be inaccurate due to low resolution imagery [5] and temporal variations in illumination and appearance [29, 79], or even subjective, especially in presence of external occlusions [38, 89] and image blur effects [68, 96]. Besides, modeling self-occluded object parts is proven to be an ambiguous task since 3D consistent keypoint annotations are needed [97]. As a consequence, supervised approaches are prone to learning suboptimal models from noisy training data.

Unsupervised keypoint detection methods can predict consistent keypoint structures [19–21, 27, 43, 98], but they lack human interpretability or may be insufficient, e.g., for editing tasks requiring detailed manipulation of object parts. Jakab

*Work was done while interning at Flawless AI et al. [28] pioneered adding interpretability with a cycle loss between unsupervised images and unpaired pose examples. However, their focus is different and the proposed cycleGAN struggles in a few-shot setting as it does not exploit paired examples. Unsupervised methods can be extended to few-shot setups either by learning a mapping that regresses detected keypoints to human-labeled annotations [73], which requires hundreds or thousands of examples, or by attaching few-shot annotated examples to the unsupervised training batch as weak supervision [51]. However, we show that neither approach produces competitive predictions when added to state-of-the-art unsupervised keypoint localization methods. Our contributions focus on making the latter approach work, building upon the unsupervised reconstruction in [20] and the skeleton formulation in [28].

Recent advances in semi-supervised keypoint localization has shown significant progress in the field. Still, most existing methods are specialized for a single object category such as faces [3,57,85] and X-rays [6,94,95,100], or require hundreds or thousands of annotated examples to achieve competitive performance [51,56,81]. Our approach, however, only needs a few dozens of examples. In an orthogonal direction, Honari et al. [24] assist keypoint localization via equivariance transforms and classification labels, though the latter are not always available. Generative image labeling has also shown great promise [69, 90, 99]. However, annotating StyleGAN-generated images is prone to artifacts and noise. Besides, generative approaches are limited to the underlying data distribution biases [53,71], thus decreasing overall keypoint localization performance.

Current limitations create the need for an approach that can leverage a smaller yet semantically consistent corpus of human labeled annotations while generalizing to a much larger unlabelled image set. This paper presents a novel formulation that learns to localize semantically consistent keypoint definitions, even for occluded regions, for various object categories with complex geometry using only a few user-labeled images. We use as input a few example-based user-labeled 2D images with predefined keypoint definitions and their linkages to learn to localize keypoints. Unlike unsupervised methods, the user-selected few-shot images act as semantic shape constraints for human-interpretable keypoint localization. To enable generalization to the target data distribution, we extend our approach via self-supervision using a larger unlabeled dataset. In addition, we introduce 3D geometry-aware constraints to model depth and uplift 2D keypoints in 3D with viewpoint consistency, thus achieving more accurate 2D localization.

Experimental results demonstrate that our proposed approach competes with or outperforms state-of-the-art methods in few-shot keypoint localization for human faces, eyes, animals, and cars using a only few user-defined semantic examples. We also show the capabilities of our keypoint localization approach on a novel data distribution, specifically the mouth interior, which has not been attempted with previous few-shot localization approaches. Thus, our novel general-purposed formulation paves the way for semantically conditional generative modeling with a few user-labeled examples. We hope it will enable a broader set of downstream applications, including fast dataset labeling, and in-the-wild modeling and tracking of complex objects, among others. Our key contributions are summarized as follows:

1. A novel formulation for few-shot 3D geometry-aware keypoint localization that works on diverse data distributions.
2. We introduce keypoint uncertainty and local 3D aware geometry constraints for better keypoint localization.
3. We adapt techniques of transformation equivariance and image reconstruction from unsupervised methods.
4. Our approach enables flexible modeling of complex deformable objects and geometric parts, such as mouth interior, faces, eyes, cars, and animals via a few user examples with consistent semantic definitions.

2. Related Work

Supervised keypoint localization Supervised methods learn keypoint localization by leveraging a large corpus of human-labeled images for standard object categories [42] or domain-specific classes, such as faces [4, 36, 89, 97], eyes [37], teeth [80], human bodies [2, 26, 29, 48], animals [41], and vehicles [58, 91], among others. Annotating images and videos is not only expensive but also prone to labeling errors [23,47,61] due to image downsampling [5], object occlusions [38,89], image blur [68,96], and harsh appearance and lighting variations especially in video datasets [29,79]. As such, supervised models render inaccurate for downstream tasks, often requiring statistical uncertainty modeling [16,40] or robust regressors [10] to achieve state-of-the-art performance. A recent line of work resorts to problem-specific high-quality synthetic datasets, mainly for human bodies [54], faces [86], eyes [52,88] and teeth [87] to produce perfect semantic annotations, even for partially occluded object parts. Despite these efforts, creating a synthetic dataset with real-world data distribution remains a challenging and very laborious task, and methods trained on them still achieve near-competitive performance [52,86,87]. On the other hand, our approach can learn a general model that preserves semantic definitions from limited annotations and user constraints while still generalizing well to an unseen distribution via self-supervision.

Semi-supervised keypoint localization We draw a line between semi-supervised keypoint localization, where hundreds or thousands of image annotations are needed, and few-shot keypoint localization, where the number is restricted to dozens. Qian et al. [57] transfer style of labeled images to augment the face appearance distribution of the training set. Dong and Yang [11] propose a teacher network to select pseudo labels generated by student networks. The methods above handle limited head pose variations. Wang et al. [81] extend the pseudo label idea to general image distribution by exploiting reinforcement learning, which can be unstable and computationally expensive. Extra information, such as classification labels [24, 76], multiview constraints [14], and video frames [56] can help in semi-supervised learning, but it is not always available for all datasets. Mathis et al. [46] fine-tune a pose estimation network [25] on hundreds of labeled images for constrained animal pose detection tasks under simple laboratory conditions. Moskvyak et al. [51] directly learn from unlabeled images, akin to our method. They impose equivariance constraints between images and keypoints such that features extracted at keypoints remain invariant to linear transformations. These semi-supervised methods still require a labeled dataset that is one or two orders of magnitude higher than that of our approach.

Few-shot keypoint localization Most existing few-shot keypoint localization methods focus on specific domains, mostly faces [3, 85] or medical X-ray images [6, 94, 95, 100]. Browatzki et al. [3] pre-train an auto-encoder on millions of faces and modify it to generate keypoints. Similarly, Thewlis et al. [72] pre-train on tens of thousands of faces and toy roboarms. Wei et al. [85] fine-tune a pre-trained face landmark detector for custom keypoint locations. In the medical domain, X-ray images often share common appearance and viewpoints. Thus, state-of-the-art methods constrain 2D keypoint deviations [6] and the features extracted at keypoint locations [94], or resort to unsupervised registration [95] to adapt the keypoints from the few-shot examples to the unlabeled images. Although these approaches show effectiveness for X-ray images, they may not be applicable to images with larger viewpoint and appearance variations. Instead of targeting a single domain, our proposed method addresses more general object distributions. Jakab et al. [28] used a CycleGAN to transfer unpaired annotations of various objects in form of skeleton edge maps across domains and showed that it applies to the few-shot scenario. We use the same skeleton edge map but use different unsupervised learning techniques to exploit labeled samples more effectively.

Unsupervised keypoint localization Reconstructing an image from keypoints [27, 98] and moving keypoints from known or estimated image transformations [73] are common approaches in unsupervised keypoint learning. For example, keypoints should be consistent across view transformations in multi-view captures [59, 60, 70] and have consistent motion transformations in videos [13, 28, 33, 50, 65]. Furthermore, transformations can be created by artificial image deformations in single static images [27, 43, 73, 98]. Another branch is to learn by synthesizing the image without assuming any pre-defined transformation [19–21]. Here we adapt the image transformation used in [43, 73] and the image reconstruction technique from [20] into our 3D-aware framework to boost our model performance.

Generative image labeling Generative models, e.g., GANs [31, 32] and diffusion models [22], with rich semantic spatial information, can be serve powerful priors for solving downstream tasks, e.g., semantic segmentation [17, 39, 74, 93, 99], video object detection [69], and salient object detection [90] from a few annotated examples. As these methods require labeling generated images that are prone to artifacts and noise, user annotations might be inaccurate, leading to inaccurate estimated labels [99]. Also, models trained on generative image-label pairs are limited to the data distribution and biases of the learned image generator [53, 71]. Our approach, however, shows better adaptation to diverse data distributions using a few manual labels.

3. Method

Our goal is to learn semantically meaningful and consistent keypoints by using only dozens of annotated examples combined with thousands of unlabeled images. Our starting point is supervised learning on the dozens of labeled examples. This step is used to define the semantic meaning of keypoints, but alone would horrendously overfit. Hence, it is augmented with self-supervision objectives that encourage meaningful and consistent detection on the additional unlabeled set. We achieve geometry-aware keypoint localization via a multi-task learning strategy that first detects keypoints with edge and uncertainty maps, and then decodes these maps with randomly masked images to synthesize photo-realistic images, as shown in Figure 2. The edge linkages are provided by the user. We train both detection and reconstruction stages in an end-to-end manner to ensure that the predicted edges and uncertainty maps derived from keypoints encode the correct semantic object shape definition to synthesize a photo-realistic object. Note that we define the uncertainty differently from previous work [64] where it refers to Gaussian shape.

In the following sections, we provide details on the different stages and our few-shot training strategy.

3.1. Keypoint and Uncertainty Detection

In this section, we introduce how we obtain the keypoints and uncertainty from the images. Given an image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, we use a ResNet with upsamplings [92] to predict K heatmaps $\mathbf{H}_i \in \mathbb{R}^{H\times W}$, and K uncertainty maps $\mathbf{V}_i \in \mathbb{R}^{H\times W}$, where $i = 1..., K$. The 2D keypoints $\mathbf{k}_i \in \mathbb{R} \in [-1, 1]^2$ are generated as the arg-softmax of the heatmap, and the uncertainty $\mathbf{v}_i$ is calculated as the sum of the map $\mathbf{V}_i$ weighted by the heatmap, as follows,

$$\mathbf{k}_i = \sum_{\mathbf{p}} w(\mathbf{p})\mathbf{p}, \quad \mathbf{v}_i = \sum_{\mathbf{p}} w(\mathbf{p})\mathbf{V}_i(\mathbf{p}), \\ \text{where } w(\mathbf{p}) = \frac{\exp(\mathbf{H}_i(\mathbf{p}))}{\sum_{\mathbf{p}} \exp(\mathbf{H}_i(\mathbf{p}))}. \tag{1}$$

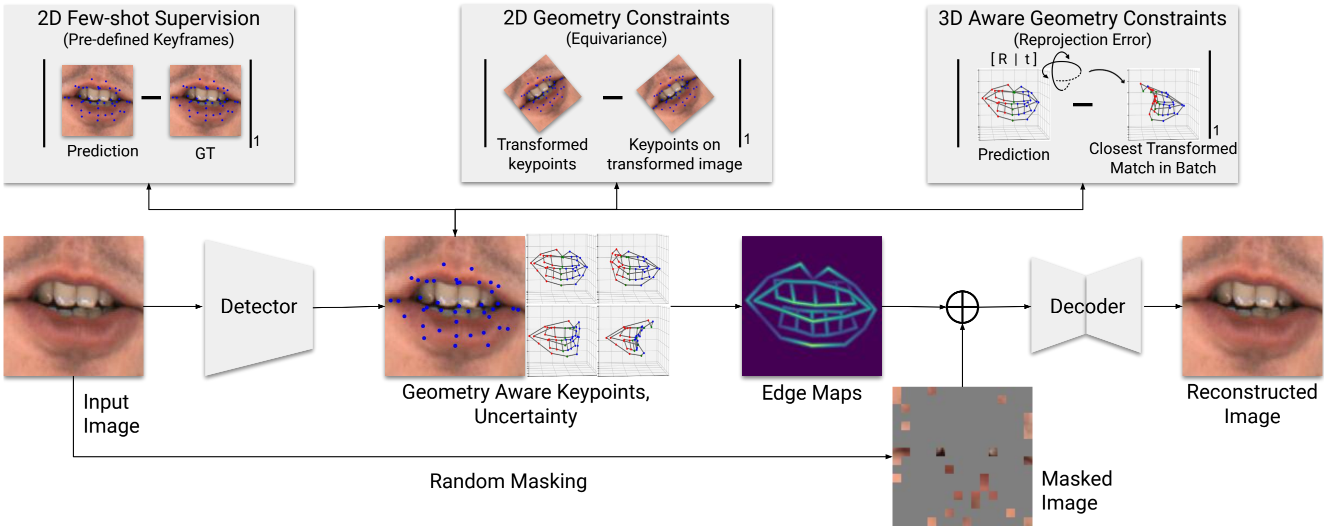

Figure 2. **Overview**. Given an image, we detect the keypoints and their uncertainty. They are used to generate an edge map, which is concatenated with a randomly masked image to reconstruct the original image. The keypoints are forced to be semantically meaningful by few-shot supervision and consistent by reconstruction. In addition, the 2D and 3D geometric constraints increase the robustness of keypoints.

Here, $\mathbf{p} \in [-1, 1]^2$ denotes the normalized pixel coordinate.

3.2. 2D Few-shot Supervision

To leverage the user-provided few-shot 2D keypoint annotations, during each training iteration, we randomly select several image-keypoint pairs, along with a batch of images without any annotations. We concatenate them and penalize deviations of the detected keypoints $\mathbf{k}'$ from the ground truth keypoints $\mathbf{k}$ on those with annotations,

$$\mathcal{L}_{\text{few\_shot}} = \frac{1}{|A|} \sum_{i \in A} \|\mathbf{k}_i - \mathbf{k}'_i\|_1, \quad (2)$$

where A is the set of annotated examples.

3.3. 2D Geometric Constraints

We force the keypoints to be equivariant to the 2D image transformations, which benefits the robustness as suggested in various unsupervised keypoint detection methods [27, 43, 73]. Let denote $\mathcal{T}$ as a 2D transformation, which is a combination of affine transformations, flipping and color jitter, and $\mathbf{k}(\mathbf{I})$ as the keypoints detected on image $\mathbf{I}$. We force the transformed keypoints $\mathcal{T}(\mathbf{k}(\mathbf{I}))$ to be close to the keypoints $\mathbf{k}(\mathcal{T}(\mathbf{I}))$ detected on the transformed image $\mathcal{T}(\mathbf{I})$,

$$\mathcal{L}_{\text{2d\_geo}} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{k}(\mathcal{T}(\mathbf{I}_i)) - \mathcal{T}(\mathbf{k}(\mathbf{I}_i))\|_1. \quad (3)$$

The keypoints that are out of image boundary after the transformation are ignored. We notice that this equivariance loss significantly harms the performance in the first few iterations during training. Therefore, we linearly increase the range of image transformation based on the number of iterations to stabilize training.

3.4. 3D Aware Geometry Constraints and Uplifting

In the domain of multi-view unsupervised 3D keypoint learning, multi-view consistency is usually used [60, 70]. We would like to use 3D consistency for better robustness but face the challenge of not having multiple views to lift to 3D.

First, to move from a 2D to a 3D keypoint representation, we extend the detector to generate K depth maps $\mathbf{D}_i \in \mathbb{R}^{H \times W}, i = 1, ..., K$ and calculate the depth as the weighted sum of the depth maps,

$$\mathbf{d}_i = \sum_{\mathbf{p}} w(\mathbf{p}) \mathbf{D}_i(\mathbf{p}), \quad (4)$$

where $w(\mathbf{p})$ is the weight defined in Equation 1. The 3D keypoints are defined as the concatenation of the 2D keypoints and the corresponding depths, $\mathbf{k}_i^{3D} = (\mathbf{k}_i, \mathbf{d}_i)$.

Second, we learn these 3D keypoints in the absence of 3D labels and multiple views by exploiting that different instances of the same object are self similar in 3D. It would be problematic to simply enforce the similarity of all keypoints on different objects as it would force different instances to be exactly the same. To address this issue, we propose to constrain their similarity separately within each part $\mathcal{P}$. Note that the entire object is also included, but as a soft constraint with a lower weight, thus enabling local deformation. Parts are pre-defined by the user, e.g., the connected components on WFLW face dataset, which are the left/right eye, left/right brow, nose, mouth, and facial contour. Specifically, for the keypoints $\mathbf{k}^{3D}(\mathbf{I})$ detected on image $\mathbf{I}$, we select another set of keypoints $\mathbf{k}^{3D}(\mathbf{I}_c)$ detected on image $\mathbf{I}_c$. For each part $\mathcal{P}$, i.e., subsets or the whole set of keypoints, we estimate a similarity transformation η between the them [77], and force the transformed keypoints $\eta(\mathbf{k}^{3D}_{\mathcal{P}}(\mathbf{I}_c))$ to be close to the keypoints $\mathbf{k}^{3D}_{\mathcal{P}}(\mathbf{I})$ of the first example,

$$\mathcal{L}_{\text{3d\_geo}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{\mathcal{P}}\frac{1}{|\mathcal{P}|}\min_{I_c}\|\mathbf{k}^{3D}_{\mathcal{P}}(\mathbf{I}_i) - \eta\left(\mathbf{k}^{3D}_{\mathcal{P}}(\mathbf{I}_c)\right)\|_1. \tag{5}$$

To prevent early overfitting, for the first 200 iterations, we randomly pair the parts in the batch. Afterwards, we pair each part with the one in the same batch that minimizes the mean L_2 distances as shown in Equation 5.

3.5. Geometry-aware Image Reconstruction

Finally, we exploit that the keypoints should be semantically meaningful enough to reconstruct the original image [20, 27, 28, 43, 98]. Specifically, we reconstruct from a largely masked image [20]. Similar to [20], we introduce an objective to reconstruct the original image from the edge map, with appearance provided by a masked image (90% of pixels removed). We take the keypoint uncertainty into consideration instead of simply assuming all keypoints are certain as in [20, 28]. Given two keypoints $\mathbf{k}_i, \mathbf{k}_j$ defined as "linked" by users, we draw a differentiable edge map $\mathbf{S}_{ij}$, where edge is a Gaussian extended along the line [20, 28, 49]. The values decrease exponentially based on the distance to the line, and are smaller for the uncertain keypoints. Formally, the edge map $\mathbf{S}_{ij}$ of keypoints $(\mathbf{k}_i, \mathbf{k}_j)$ is defined as

$$\mathbf{S}_{ij}(\mathbf{p}) = \exp\left(v_{ij}(\mathbf{p})d^2_{ij}(\mathbf{p})/\sigma^2\right), \tag{6}$$

where σ is a learnable parameter controlling the thickness of the edge, $d_{ij}(\mathbf{p})$ is the L_2 distance between the pixel $\mathbf{p}$ and the edge drawn by keypoints $\mathbf{k}_i$ and $\mathbf{k}_j$, and $v_{ij}(\mathbf{p})$ is the uncertainty propagated to pixel $\mathbf{p}$ along the edge $\mathbf{k}_i$ to $\mathbf{k}_j$,

$$v_{ij}(\mathbf{p}) = \begin{cases} \text{sigmoid}(\mathbf{v}_i) & \text{if } t \le 0, \\ \text{sigmoid}((1-t)\mathbf{v}_i + t\mathbf{v}_j) & \text{if } 0 < t < 1, \\ \text{sigmoid}(\mathbf{v}_j) & \text{if } t \ge 1, \end{cases}$$
$$\text{where} \quad t = \frac{(\mathbf{p}-\mathbf{k}_i)\cdot(\mathbf{k}_j-\mathbf{k}_i)}{\|\mathbf{k}_i-\mathbf{k}_j\|_2^2} \tag{7}$$

is the normalized distance between $\mathbf{k}_i$ and the projection of $\mathbf{p}$ onto the edge. We assign a learnable weight α to edges, which is enforced to be positive by SoftPlus [12]. This weight is learned during training and shared across all edges and all object instances in a dataset. Finally, we take the maximum at each pixel of the heatmaps to obtain the edge map $\mathbf{S} \in \mathbb{R}^{H\times W}$,

$$\mathbf{S}(\mathbf{p}) = \alpha \max_{ij} \mathbf{S}_{ij}(\mathbf{p}). \tag{8}$$

Taking the maximum at each pixel avoids the entanglement of the uncertainty and the convolution kernel weights [20].

The edge map is concatenated with the masked image and fed into a UNet [62] to reconstruct the original image. We minimize the L_1 loss and ViT perceptual loss [75] between the reconstructed images $\mathbf{I}'$ and the original images $\mathbf{I}$,

$$\mathcal{L}_{\text{recon}} = \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{I}_i - \mathbf{I}'_i\|_1 + \|\Gamma(\mathbf{I}_i) - \Gamma(\mathbf{I}'_i)\|_1, \tag{9}$$

where Γ is the feature extractor.

3.6. Coefficients and Few-shot Examples Chosen

The coefficient for $\mathcal{L}_{\text{few\_shot}}, \mathcal{L}_{\text{recon}}, \mathcal{L}_{\text{2d\_geo}}, \mathcal{L}_{\text{3d\_geo}}$ are (1, 1, 1, 0.1), which is decided by the validation sets. There is no other hyperparameters. Note that the validation sets are not used to choose the best model during training but only used to find the loss coefficients, which are shared by all datasets. We believe that using the validation sets for model selection in practice is not available in few-shot learning. The few-shot examples are chosen by the centers of k-means clustering on the features of the 3nd last layer of VGG [67]. More implementation details can be found in the Supplement A.

4. Results

We test our model on 6 diverse datasets, including rigid, soft, and articulated objects, which may contain various appearance and severe occlusion. Notably the mouth interior is extremely challenging due to the large occlusions and had not been attempted before. We compare with three baselines methods: supervised [92], semi-supervised [51], and unsupervised [20]. We adapt the latter to few-shot learning using the strategy in Section 3.2. Unless otherwise stated, all results are obtained by 10-shot learning except Tigers where 20 examples are used. See Supplement B for comparisons with DatasetGAN [99].

4.1. Datasets

MEAD Part0 [82] contains high resolution audio-visual clips of 12 actors. We use the first actor and crop around the mouth. We label 10 images with the lip landmarks from [16] and manual annotations on 4 top and 5 bottom teeth.

SynthesEyes [88] contains 11382 synthesized eye images from 5 male and 5 female subjects. We use the first 4 males and 4 females as our training set and the rest for testing.

CUB [78] contains 200 categories of birds. We use the first 100 for training (5864 images), the next 50 for validation (2958 images), and the last 50 for testing (2966 images). This dataset is to quantitatively evaluate the ability of detecting keypoints on objects of highly various appearance.

CarFusion [58] contains various car images captured in Pittsburgh, PA. We use Fifth Street Part 2 (2794 images) and Part 1 (1597 images) as training and testing sets.

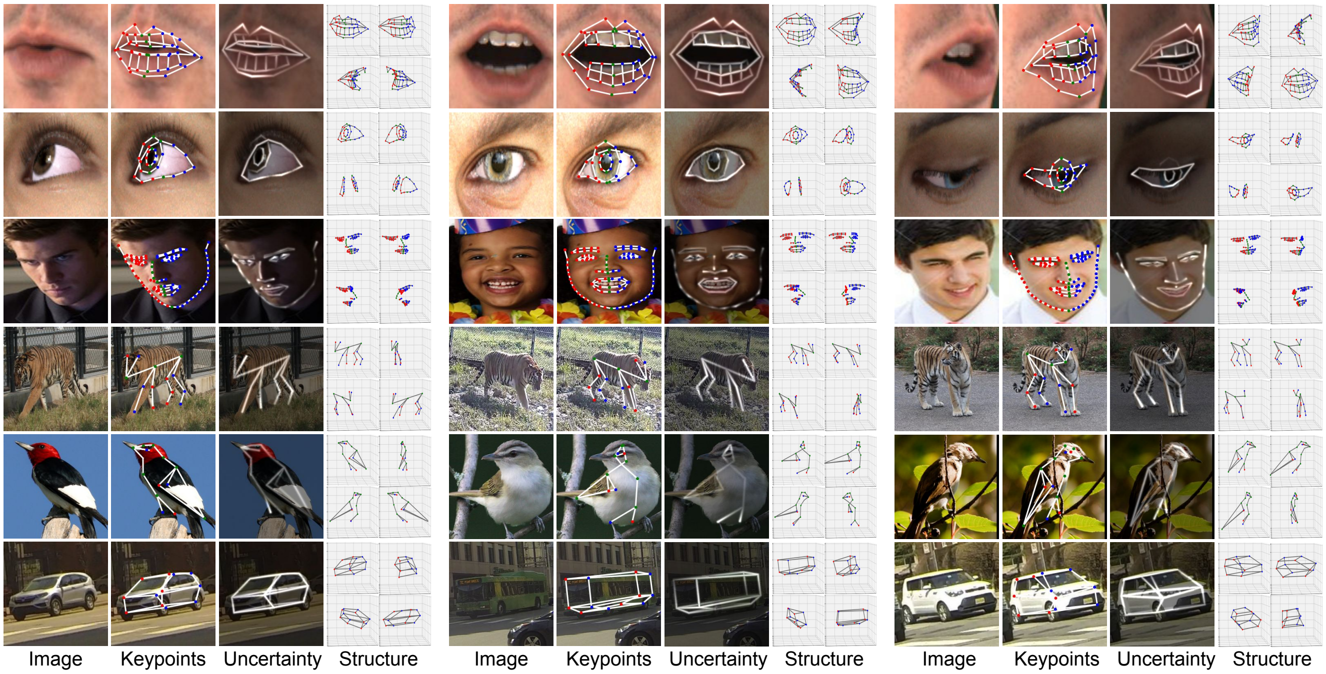

Figure 3. **Qualitative Results**. With only few shots, the model learns semantically consistent and meaningful keypoints.

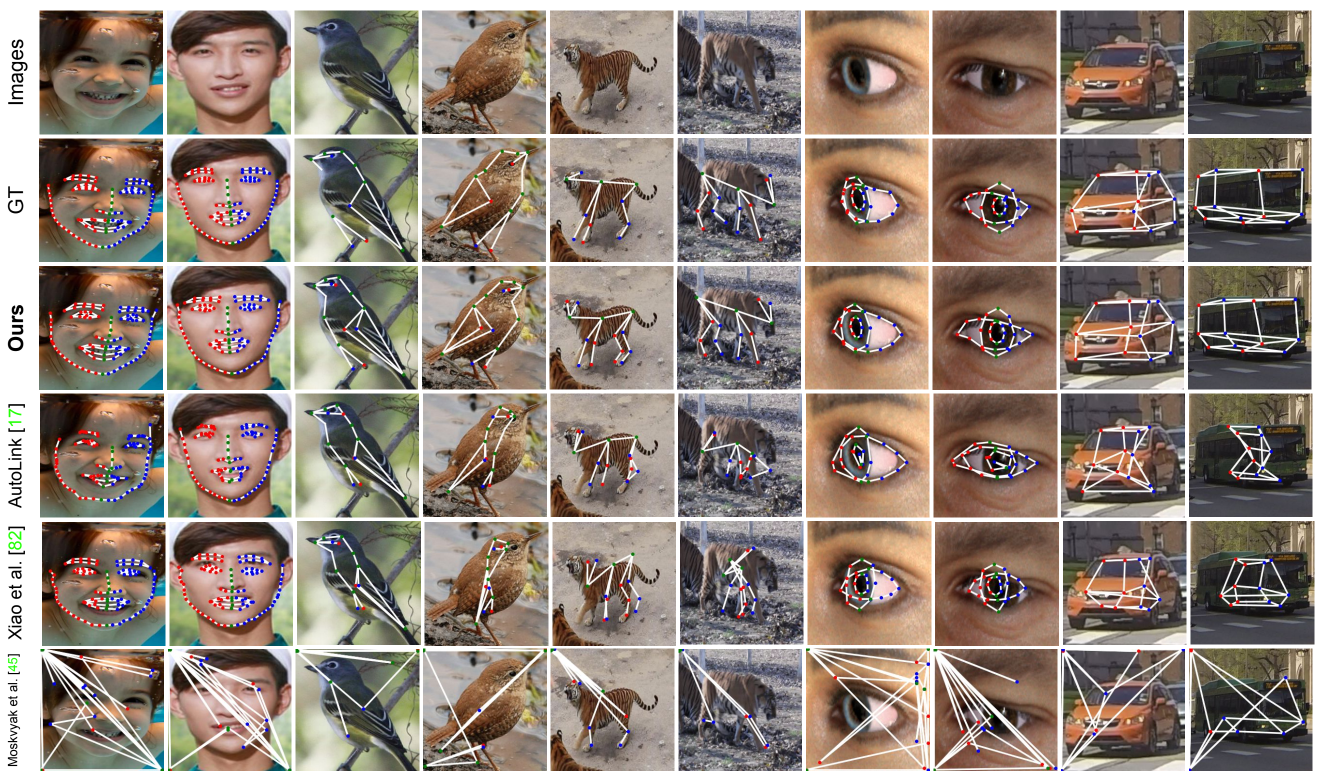

Figure 4. **Qualitative Comparison**. We qualitatively compare our results with the baselines and the ground truth. The invisible keypoints are not shown in the ground truth if they are not provided. Our performance is significantly better than the other baselines. In the difficult cases, such as tigers and cars, our model still generates good shapes while others fail.

WFLW [89] contains 10k faces with a 7500/2500 train/test split. This dataset is more challenging than most face datasets as images have a significant portion of occlusions, make-up, and extreme poses.

ARTW [41] contains 5159 tiger images captured from multiple wild zoos in unconstrained settings, with a 3610/516/1033 train/val/test split.

SynthesisAI/Faces [1] contains 10k images of 100 diverse identities with ground truth 3D facial landmarks, including the occluded ones. We split the dataset into 7500 images for training and 2500 images for testing. We use this dataset to quantitatively measure how good the 3D landmarks that our 3D geometry constraint yields are.

4.2. Qualitative Comparisons

We show the qualitative results in Figure 3 and comparisons in Figure 4. When the number of annotations are limited, the baselines tend to overfit and fail to detect on the difficult cases. For example, only our model successfully detects the keypoints of the bus in the last column of Figure 4.

4.3. Quantitative Comparisons

We quantitatively compare our model on 5 benchmarks: WFLW, SynthesEyes, CUB, ATRW, and CarFusion.

Evaluation Metrics are Normalized L_2 error (NME) for SynthesEyes and WFLW, and percentage of correct keypoints (PCK) for CUB, ATRW, and CarFusion. The NME is normalized by distance of eye corners and inter-ocular distances for SynthesEyes and WFLW, respectively. The PCK@0.1 is the ratio of keypoints within a range of 10% largest bounding box length centered by the GT keypoints. To benefit future research, we also report the other metrics on these datasets in Supplemental F.

Table 1 shows that our model significantly outperforms other methods when only 10-50 annotated examples are available. Furthermore, our model is more robust to the number of annotated examples than other methods. For instance, on ATRW, the accuracy of the baselines decrease (58%/61%/40%) if only 50 out of 3610 examples are used, while our performance only decreases 16%. Supplemental D provides comparisons with unsupervised methods [20, 28] on their commonly used datasets: 300W [63] and H36M [26].

NME (%) on WFLW dataset ↓

Method	Training set size						
	1	10	20	50	5%	20%	100%
SA [24]	-	-	-	-	-	6.00†	**4.39**
Xiao et al. [92]	43.0	21.9	19.3	17.6	10.6	7.08	5.62
Moskvyak et al. [51]	137	133	76.6	21.9	10.27	6.84	6.65
AutoLink (few) [20]	14.9	13.5	13.3	11.2	7.68	7.31	6.35
3FabRec [3]	15.8†	9.66†	-	8.39†	7.68†	6.51†	5.62
ours	**12.4**	**9.19**	**8.62**	**7.90**	**6.22**	**5.61**	5.38

NME (%) on SynthesEyes dataset ↓

Method	Training set size						
	1	10	20	50	5%	20%	100%
Xiao et al. [92]	25.1	15.9	10.6	8.11	4.07	**3.12**	2.65
Moskvyak et al. [51]	91.1	86.8	45.8	18.5	4.52	3.24	**2.49**
AutoLink (few) [20]	26.4	14.2	8.78	7.80	4.28	3.32	2.86
ours	**24.0**	**6.93**	**6.83**	**5.50**	**3.69**	3.13	2.96

PCK@0.1 (%) on CUB-200-2011 dataset ↑

Method	Training set size						
	1	10	20	50	5%	20%	100%
Xiao et al. [92]	6.01	31.2	35.3	42.8	60.5	73.9	90.5
Moskvyak et al. [51]	7.38	20.9	28.3	63.2	**91.1**†	**92.4**†	**93.8**
AutoLink (few) [20]	**26.2**	35.1	41.2	51.8	67.2	75.9	87.6
ours	16.3	**70.7**	**73.1**	**75.1**	84.2	88.3	90.1

PCK@0.1 (%) on ATRW dataset ↑

Method	Training set size						
	1	10	20	50	5%	20%	100%
Xiao et al. [92]	13.2	21.5	22.5	23.9	51.6	86.1	96.1
Moskvyak et al. [51]	3.38	17.9	27.5	57.1	92.6†	94.5†	95.3
AutoLink (few) [20]	**20.8**	22.1	33.8	37.4	77.8	89.4	**97.1**
ours	14.4	**36.3**	**78.8**	**81.4**	**92.7**	**96.5**	96.9

PCK@0.1 (%) on CarFusion dataset ↑

Method	Training set size						
	1	10	20	50	5%	20%	100%
Xiao et al. [92]	11.7	23.4	30.0	37.5	51.1	77.9	89.7
Moskvyak et al. [51]	3.84	5.38	22.4	34.7	64.7	87.3	92.5
AutoLink (few) [20]	**14.5**	31.9	42.3	62.2	69.9	83.5	90.8
ours	13.8	**66.7**	**68.8**	**75.5**	**81.6**	**93.5**	**93.8**

Table 1. **Quantitative Comparison**. In the few-shot scenario, where only 10-50 annotated examples are available, our model significantly outperforms the baselines. The sign † means the number is reported in another set of examples used in their papers.

Methods	Training set size						
	1	10	20	50	5%	20%	100%
Xiao et al. [92]	36.5	22.5	22.1	17.8	8.33	**5.37**	**4.19**
Full	**15.6**	**11.2**	**9.25**	**8.46**	**7.07**	6.44	5.99

Table 2. **NME(%) on SynthesisAI 3D Faces.** We compare our approach with a supervised baseline. Note that the latter uses ground truth 3D keypoints, while ours only needs 2D keypoints.

5. Analysis

We test the effects and the necessity of our designed modules. Quantitative comparisons are implemented on WFLW and ATRW, with the same settings and the same few-shot examples as in Section 4.

Image Reconstruction. Table 3 shows that without the image Reconstruction constraint, the error increases dramatically for low numbers of annotated examples. It is believed that reconstructing from edge maps aligns the edges with the object edges [20], which stabilize the keypoints in the few-shot learning. Note that with more than 20% annotated examples, the image reconstruction harms the performance, probably because the reconstruction task drifts the gradient from accurate keypoint supervision gradient.

Geometric Constraint. While both 2D and 3D geometric constraints increase accuracy, their insights differ slightly.

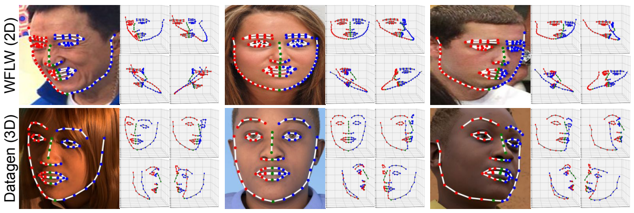

Figure 5. **Detection on 2D and 3D-projected Face Landmarks**. If the model is trained on the 2D face dataset WFLW, the learned jaw landmarks are not 3D consistent. However, if it is trained on the synthetic 3D-projected landmarks, the jaw contours are 3D consistent, though the projection is still on the 2D facial contour.

The 2D constraint is at the image level, which is accurate. It works as data augmentation on learned keypoints, which is expected to increase the robustness. Interestingly, the variant without 2D geometric constraint has better performance on Tigers (10-shots). We believe it is due to the symmetric shape of the tigers, as explained in the limitations. On the other hand, the 3D constraint is an approximation and enforces similarity between two different examples. This constraint is expected to prevent the model from generating extreme outliers. Its effects on articulated objects, e.g., tigers, are not as significant as they are on soft and rigid objects, e.g., faces.

Uncertainty. Uncertainty models occluded and ambiguous edges in image reconstruction by making the uncertain edges lighter. Without it, the model may be confused in the reconstruction stage about whether occluded edges should be drawn. As a result, it harms performance. Figure 6(a) shows an example of a largely occluded face. The model without uncertainty gives an average face. On ATRW, the difference is less obvious as the keypoints are visible in most cases.

How good are the 3D keypoints? We train on the SynthesisAI/Faces [1] using only 2D landmarks and compare the 3D landmarks quality with the supervised baseline [92], where depth is additionally learned. We evaluate the learned 3D landmarks qualitatively in Figure 5 and quantitatively in Table 2. Figure 5 also shows the difference between the jaw landmarks learned from a 2D and 3D-projected facial landmark dataset. The signal provided by the latter gives a better 3D jaw shape since the few-shot landmarks do not follow visible image boundaries as those in the 2D facial dataset.

Few-shot Example Selection. We tested replacing the KMeans with random selection. Table 3 reports the average error over 3 runs. When only dozens of annotated examples are available, it is important to pick the most representative ones, especially for articulated objects, such as ATRW tigers.

Limitations. There are two limitations especially when the annotated dataset is very small. First, if the object's keypoints are highly symmetric, there may be left-right or front-back ambiguity. Second, if the object is highly articulated, e.g., humans in LSP dataset [30], the estimated keypoints are less accurate. Figure 6(b) illustrates both cases.

NME (%) on WFLW dataset ↓

Variants	Training set size						
	1	10	20	50	5%	20%	100%
- image reconstruction	27.4	24.8	20.2	12.4	6.49	**5.32**	**4.69**
- 2D geometry	16.7	14.2	13.7	12.6	9.24	7.81	6.83
- 3D geometry	15.5	12.4	12.1	11.3	7.67	6.47	6.37
- uncertainty	13.5	12.6	11.8	10.4	7.28	6.75	6.42
- kmeans selection	24.2	14.5	12.6	10.8	6.81	5.54	-
Full	**12.4**	**9.19**	**8.62**	**7.90**	**6.22**	5.61	5.38

PCK@0.1 (%) on ATRW dataset ↑

Variants	Training set size						
	1	10	20	50	5%	20%	100%
- image reconstruction	20.2	22.1	22.8	56.5	91.0	**96.8**	**97.6**
- 2D geometry	24.7	**40.6**	39.3	39.6	87.4	93.7	96.1
- 3D geometry	16.7	36.2	75.8	81.2	92.1	96.1	96.4
- uncertainty	20.8	31.9	78.7	81.2	92.3	95.9	95.4
- kmeans selection	8.2	10.1	15.1	40.7	90.9	96.1	-
Full	14.4	36.3	**78.8**	**81.4**	**92.7**	96.5	96.9

Table 3. **Ablation Tests on WFLW and ATRW.** Each of our design choices plays an important role in the few-shot scenario.

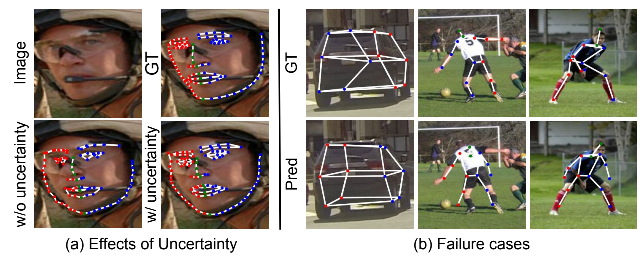

Figure 6. **(a) Effects of Uncertainty**. Modeling uncertainty improves keypoint localization on objects with occlusion (see for instance the nose and eyebrows). **(b) Failure Cases**. The model fails if the objects are highly symmetric or articulated.

However, neither problem is observed when we increase the dataset to hundreds of annotations (instead of dozens).

6. Conclusion & Discussion

We presented a few-shot keypoint localization method that is formed by combining keypoint detection with uncertainty, 2D/3D geometric constraints, and image reconstruction. These components prevent the detector from overfitting to the few-shot examples and utilize unlabelled images. Our experiment results demonstrate that, with only dozens of annotations, our model works on various datasets, including rigid, soft, articulated objects, and even the very difficult mouth interior which has not been tried before. In the few-shot scenario, our model significantly outperforms the baselines and works on those datasets where others fail with 10 or 20 annotated examples. It opens the path for conditional generative modeling and image editing with a few annotated examples. Our future work will focus on leveraging 3D-aware image synthesis for better generalization to extreme poses, solving the symmetry problem, and testing on broader object categories.

References

[1] Synthesis AI. Face api dataset. https://github.com/Synthesis-AI-Dev/face_api_dataset, 2022. [Online; accessed 21-Sep-2022]. 7, 8

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter V. Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693. IEEE Computer Society, 2014. 2

[3] Björn Browatzki and Christian Wallraven. 3fabrec: Fast few-shot face alignment by reconstruction. In *CVPR*, pages 6109–6119. Computer Vision Foundation / IEEE, 2020. 2, 3, 7

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

[5] Prashanth Chandran, Derek Bradley, Markus H. Gross, and Thabo Beeler. Attention-driven cropping for very high resolution facial landmark detection. In *CVPR*, pages 5860–5869. Computer Vision Foundation / IEEE, 2020. 1, 2

[6] Runnan Chen, Yuexin Ma, Lingjie Liu, Nenglun Chen, Zhiming Cui, Guodong Wei, and Wenping Wang. Semi-supervised anatomical landmark detection via shape-regulated self-training. *Neurocomputing*, 471:335–345, 2022. 2, 3

[7] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. Potion: Pose motion representation for action recognition. In *CVPR*, pages 7024–7033. Computer Vision Foundation / IEEE, 2018. 1

[8] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. *Workshop on Stat. Learn. Comput. Vis., ECCV*, Vol. 1:1–22, 2004. 1

[9] Debayan Deb and Anil K. Jain. Look locally infer globally: A generalizable face anti-spoofing approach. *IEEE Trans. Inf. Forensics Secur.*, 16:1143–1157, 2021. 1

[10] Piotr Dollár, Peter Welinder, and Pietro Perona. Cascaded pose regression. In *CVPR*, pages 1078–1085. IEEE Computer Society, 2010. 2

[11] Xuanyi Dong and Yi Yang. Teacher supervises students how to learn from partially labeled images for facial landmark detection. In *ICCV*, pages 783–792. IEEE, 2019. 2

[12] Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. In *NeurIPS*, pages 472–478. MIT Press, 2000. 5

[13] Aysegul Dundar, Kevin J Shih, Animesh Garg, Robert Pottorf, Anrew Tao, and Bryan Catanzaro. Unsupervised disentanglement of pose, appearance and background from images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3883–3894, 2021. 3

[14] Qi Feng, Kun He, He Wen, Cem Keskin, and Yuting Ye. Active learning with pseudo-labels for multi-view 3d pose estimation. *arXiv preprint arXiv:2112.13709*, 2021. 3

[15] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, volume 11218 of *Lecture Notes in Computer Science*, pages 557–574. Springer, 2018. 1

[16] David Ferman and Gaurav Bharaj. Multi-domain multi-definition landmark localization for small datasets. *CoRR*, abs/2203.10358, 2022. 2, 5

[17] Mengya Han, Heliang Zheng, Chaoyue Wang, Yong Luo, Han Hu, and Bo Du. Leveraging gan priors for few-shot part segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1339–1347, 2022. 3

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *ICCV*, pages 2980–2988. IEEE Computer Society, 2017. 1

[19] Xingzhe He, Bastian Wandt, and Helge Rhodin. Latentkeypointgan: Controlling gans via latent keypoints. *arXiv preprint arXiv:2103.15812*, 2021. 1, 3

[20] Xingzhe He, Bastian Wandt, and Helge Rhodin. Autolink: Self-supervised learning of human skeletons and object outlines by linking keypoints. In *NeurIPS*, 2022. 1, 2, 3, 5, 7, 13, 14, 15

[21] Xingzhe He, Bastian Wandt, and Helge Rhodin. Ganseg: Learning to segment by unsupervised hierarchical image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1225–1235, 2022. 1, 3

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 3

[23] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *ECCV*, volume 7574 of *Lecture Notes in Computer Science*, pages 340–353. Springer, 2012. 2

[24] Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher J. Pal, and Jan Kautz. Improving landmark localization with semi-supervised learning. In *CVPR*, pages 1546–1555. Computer Vision Foundation / IEEE, 2018. 2, 3, 7

[25] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, volume 9910 of *Lecture Notes in Computer Science*, pages 34–50. Springer, 2016. 3

[26] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE TPAMI*, 36(7):1325–1339, 2014. 2, 7, 13

[27] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. *Advances in neural information processing systems*, 31, 2018. 1, 3, 4, 5

[28] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8787–8797, 2020. 2, 3, 5, 7, 14

[29] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *ECCV*, volume 12354 of *Lecture Notes in Computer Science*, pages 196–214. Springer, 2020. 1, 2

[30] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, pages 12.1–12.11. BMVA Press, 2010. doi:10.5244/C.24.12. 8, 13

[31] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3, 13

[32] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 3

[33] Yunji Kim, Seonghyeon Nam, In Cho, and Seon Joo Kim. Unsupervised keypoint learning for guiding class-conditional video prediction. In *Advances in Neural Information Processing Systems*, pages 3814–3824, 2019. 3

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

[35] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 13

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

[37] Rakshit Sunil Kothari, Zhizhuo Yang, Christopher Kanan, Reynold Bailey, Jeff B. Pelz, and Gabriel J. Diaz. Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *CoRR*, abs/1905.13146, 2019. 2

[38] Abhinav Kumar, Tim K. Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *CVPR*, pages 8233–8243. Computer Vision Foundation / IEEE, 2020. 1, 2

[39] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8300–8311, 2021. 3

[40] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *ICCV*, pages 11005–11014. IEEE, 2021. 2

[41] Shuyuan Li, Jianguo Li, Hanlin Tang, Rui Qian, and Weiyao Lin. ATRW: A benchmark for amur tiger re-identification in the wild. In *ICM*, pages 2590–2598. ACM, 2020. 2, 7

[42] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014. 2

[43] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Bjorn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10955–10964, 2019. 1, 3, 4, 5

[44] Yuhang Lu, Weijian Li, Kang Zheng, Yirui Wang, Adam P. Harrison, Chihung Lin, Song Wang, Jing Xiao, Le Lu, Chang-Fu Kuo, and Shun Miao. Learning to segment anatomical structures accurately from one exemplar. In *MICCAI*, volume 12261 of *Lecture Notes in Computer Science*, pages 678–688. Springer, 2020. 1

[45] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *CVPR*, pages 99–108. Computer Vision Foundation / IEEE, 2018. 1

[46] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, M. Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21:1281–1289, 2018. 3

[47] Alexander Mathis, Steffen Schneider, Jessy Lauer, and Mackenzie W. Mathis. A primer on motion capture with deep learning: Principles, pitfalls, and perspectives. *Neuron*, 108(1):44–65, 2020. 2

[48] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved CNN supervision. In *3DV*, pages 506–516. IEEE Computer Society, 2017. 2

[49] Daniela Mihai and Jonathon S. Hare. Differentiable drawing and sketching. *CoRR*, abs/2103.16194, 2021. 5

[50] Matthias Minderer, Chen Sun, Ruben Villegas, Forrester Cole, Kevin P Murphy, and Honglak Lee. Unsupervised learning of object structure and dynamics from videos. In *Advances in Neural Information Processing Systems*, pages 92–102, 2019. 3

[51] Olga Moskvyak, Frédéric Maire, Feras Dayoub, and Mahsa Baktashmotlagh. Semi-supervised keypoint localization. In *ICLR*. OpenReview.net, 2021. 2, 3, 5, 7, 13, 14, 15

[52] Nitinraj Nair, Rakshit Sunil Kothari, Aayush K. Chaudhary, Zhizhuo Yang, Gabriel J. Diaz, Jeff B. Pelz, and Reynold J. Bailey. Rit-eyes: Rendering of near-eye images for eye-tracking applications. In *ACM SAP*, pages 5:1–5:9. ACM, 2020. 2

[53] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *CVPR*, pages 13493–13503. IEEE, 2022. 2, 3

[54] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: avatars in geography optimized for regression analysis. In *CVPR*, pages 13468–13478. Computer Vision Foundation / IEEE, 2021. 2

[55] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas,

and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, pages 10975–10985. Computer Vision Foundation / IEEE, 2019. 1

[56] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019. 2, 3

[57] 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*, pages 10153–10163, 2019. 2

[58] N. Dinesh Reddy, Minh Vo, and Srinivasa G. Narasimhan. Carfusion: Combining point tracking and part detection for dynamic 3d reconstruction of vehicles. In *CVPR*, pages 1906–1915. Computer Vision Foundation / IEEE, 2018. 2, 5

[59] Helge Rhodin, Victor Constantin, Isinsu Katircioglu, Mathieu Salzmann, and Pascal Fua. Neural scene decomposition for multi-person motion capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[60] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 750–767, 2018. 3, 4

[61] Matteo Ruggero Ronchi and Pietro Perona. Benchmarking and error diagnosis in multi-instance pose estimation. In *ICCV*, pages 369–378. IEEE Computer Society, 2017. 2

[62] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015. 5

[63] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013. 7, 13

[64] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *CVPR*, pages 2377–2386. Computer Vision Foundation / IEEE, 2019. 1, 3

[65] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, pages 7135–7145, 2019. 1, 3

[66] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. In *CVPR*, pages 3408–3416. Computer Vision Foundation / IEEE, 2018. 1

[67] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5, 13

[68] Keqiang Sun, Wayne Wu, Tinghao Liu, Shuo Yang, Quan Wang, Qiang Zhou, Zuochang Ye, and Chen Qian. FAB: A robust facial landmark detection framework for motion-blurred videos. In *ICCV*, pages 5461–5470. IEEE, 2019. 1, 2

[69] Vadim Sushko, Dan Zhang, Juergen Gall, and Anna Khoreva. One-shot synthesis of images and segmentation masks. *arXiv preprint arXiv:2209.07547*, 2022. 2, 3

[70] Supasorn Suwajanakorn, Noah Snavely, Jonathan J Tompson, and Mohammad Norouzi. Discovery of latent 3d keypoints via end-to-end geometric reasoning. In *Advances in neural information processing systems*, pages 2059–2070, 2018. 3, 4

[71] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *CVPR*, pages 6141–6150. Computer Vision Foundation / IEEE, 2020. 2, 3

[72] James Thewlis, Samuel Albanie, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of landmarks by descriptor vector exchange. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6361–6371, 2019. 3

[73] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *ICCV*, pages 3229–3238. IEEE Computer Society, 2017. 2, 3, 4, 14

[74] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. Repurposing gans for one-shot semantic part segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4475–4485, 2021. 3

[75] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022. 5

[76] Norimichi Ukita and Yusuke Uematsu. Semi- and weakly-supervised human pose estimation. *Comput. Vis. Image Underst.*, 170:67–78, 2018. 3

[77] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE TPAMI*, 13(4):376–380, 1991. 5

[78] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd birds 200. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 5

[79] Bastian Wandt, Marco Rudolph, Petrissa Zell, Helge Rhodin, and Bodo Rosenhahn. Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In *CVPR*, pages 13294–13304. Computer Vision Foundation / IEEE, 2021. 1, 2

[80] Ching-Wei Wang, Cheng-Ta Huang, Jia-Hong Lee, Chung-Hsing Li, Sheng-Wei Chang, Ming-Jhih Siao, Tat-Ming Lai, Bulat Ibragimov, Tomaz Vrtovec, Olaf Ronneberger, Philipp Fischer, Timothy F. Cootes, and Claudia Lindner. A benchmark for comparison of dental radiography analysis algorithms. *Medical Image Anal.*, 31:63–76, 2016. 1, 2

[81] Can Wang, Sheng Jin, Yingda Guan, Wentao Liu, Chen Qian, Ping Luo, and Wanli Ouyang. Pseudo-labeled auto-curriculum learning for semi-supervised keypoint localization. In *ICLR*. OpenReview.net, 2022. 2, 3

[82] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change

Loy. MEAD: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, volume 12366 of *Lecture Notes in Computer Science*, pages 700–717. Springer, 2020. 5

[83] Yirui Wang, Le Lu, Chi-Tung Cheng, Dakai Jin, Adam P. Harrison, Jing Xiao, Chien-Hung Liao, and Shun Miao. Weakly supervised universal fracture detection in pelvic x-rays. In *MICCAI*, volume 11769 of *Lecture Notes in Computer Science*, pages 459–467. Springer, 2019. 1

[84] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, pages 4724–4732. IEEE Computer Society, 2016. 1

[85] Zhen Wei, Bingkun Liu, Weinong Wang, and Yu-Wing Tai. Few-shot model adaptation for customized facial landmark detection, segmentation, stylization and shadow removal. *CoRR*, abs/2104.09457, 2021. 2, 3

[86] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J. Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *ICCV*, pages 3661–3671. IEEE, 2021. 2

[87] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljevic, Daniel Wilde, Stephan J. Garbin, Toby Sharp, Ivan Stojiljkovic, Tom Cashman, and Julien Valentin. 3d face reconstruction with dense landmarks. *CoRR*, abs/2204.02776, 2022. 1, 2

[88] Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. Rendering of eyes for eye-shape registration and gaze estimation. In *ICCV*, pages 3756–3764. IEEE Computer Society, 2015. 2, 5

[89] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, pages 2129–2138. Computer Vision Foundation / IEEE, 2018. 1, 2, 7

[90] Zhenyu Wu, Lin Wang, Wei Wang, Tengfei Shi, Chenglizhao Chen, Aimin Hao, and Shuo Li. Synthetic data supervised salient object detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5557–5565, 2022. 2, 3

[91] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond PASCAL: A benchmark for 3d object detection in the wild. In *WACV*, pages 75–82. IEEE Computer Society, 2014. 2

[92] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, volume 11210 of *Lecture Notes in Computer Science*, pages 472–487. Springer, 2018. 3, 5, 7, 8, 13, 14, 15

[93] Yu Yang, Xiaotian Cheng, Hakan Bilen, and Xiangyang Ji. Learning to annotate part segmentation with gradient matching. In *International Conference on Learning Representations*, 2022. 3

[94] Qingsong Yao, Quan Quan, Li Xiao, and S Kevin Zhou. One-shot medical landmark detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 177–188. Springer, 2021. 2, 3

[95] Zihao Yin, Ping Gong, Chunyu Wang, Yizhou Yu, and Yizhou Wang. One-shot medical landmark localization by edge-guided transform and noisy landmark refinement. *arXiv preprint arXiv:2208.00453*, 2022. 2, 3

[96] Chao Zhang, Xuequan Lu, and Takuya Akashi. Blur-countering keypoint detection via eigenvalue asymmetry. *IEEE Access*, 8:159077–159088, 2020. 1, 2

[97] Hongwen Zhang, Qi Li, and Zhenan Sun. Joint voxel and coordinate regression for accurate 3d facial landmark localization. In *ICPR*, pages 2202–2208. IEEE Computer Society, 2018. 1, 2

[98] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2694–2703, 2018. 1, 3, 5, 14

[99] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *CVPR*, pages 10145–10155. Computer Vision Foundation / IEEE, 2021. 2, 3, 5, 13

[100] Xiao-Yun Zhou, Bolin Lai, Weijian Li, Yirui Wang, Kang Zheng, Fakai Wang, Chihung Lin, Le Lu, Lingyun Huang, Mei Han, Guotong Xie, Jing Xiao, Chang-Fu Kuo, Adam P. Harrison, and Shun Miao. Scalable semi-supervised landmark localization for x-ray images using few-shot deep adaptive graph. In *DGM4MICCAI*, volume 13003 of *Lecture Notes in Computer Science*, pages 145–153. Springer, 2021. 2, 3