Few-shot Geometry-Aware Keypoint Localization
- Supplementary Material -

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

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

A. Implementation details

Optimization All images are resized to 128×128. We use the Adam optimizer [35] with a learning rate of 10^{-4} with $\beta_1 = 0.9$, $\beta_2 = 0.99$. The batch size is 16 for unlabeled images and $\min(16, \text{n_examples})$ for few-shot examples. We train for 20k iterations. The gradients are stopped in the similarity transformation estimation in the 3D geometric constraint so that the shape instead of transformation itself would be optimized. The ViT perceptual loss is based on the last attention keys and the the global context vector. For the reconstruction, we divide the image into 16×16 patches and randomly mask 90%. The random seeds for all packages are fixed to 0.

Formula of $\mathbf{d}_{ij}$ The distance $\mathbf{d}_{ij}$ from a pixel $\mathbf{p}$ to an edge drawn by keypoints $\mathbf{k}_i$ and $\mathbf{k}_j$ in Equation 6 is

$$d_{ij}(\mathbf{p}) = \begin{cases} \|\mathbf{p} - \mathbf{k}_i\|_2 & \text{if } t \leq 0, \\ \|\mathbf{p} - ((1-t)\mathbf{k}_i + t\mathbf{k}_j)\|_2 & \text{if } 0 < t < 1, \\ \|\mathbf{p} - \mathbf{k}_j\|_2 & \text{if } t \geq 1, \end{cases}$$
$$\text{where} \quad t = \frac{(\mathbf{p} - \mathbf{k}_i) \cdot (\mathbf{k}_j - \mathbf{k}_i)}{\|\mathbf{k}_i - \mathbf{k}_j\|_2^2} \tag{10}$$

is the normalized distance between $\mathbf{k}_i$ and the projected $\mathbf{p}$ onto the edge as in 7.

Formulation of σ and α The thickness σ in Equation 6 is formulated as

$$\sigma^2 = 1/1000 \exp \theta, \tag{11}$$

where θ is learnable and initialized to 1.

The edge map weight α in Equation 8 is formulated as

$$\alpha = \text{SoftPlus}(\gamma), \tag{12}$$

where γ is learnable and initialized to -4.

*Work was done while interning at Flawless AI

2D geometric constraint The image transformation in 2D geometric constraint is a combination of random rotation ($-60 \sim 60$), translation ($-10\% \sim 10\%$), scaling ($0.9 \sim 1$), flipping (p=0.5), and color jitter (brightness, contrast, saturation, hue from -50% to 50%). The augmentation ranges are multiplied by 0 and 1 at iteration 0 and 20k, respectively. Note that this coefficient increases linearly from 0 to 1 during training.

Facial landmark smoothing On WFLW and SynthesisAI/Faces, we encourage the cosine of the angle of the two neighboring landmarks to be 0. This penalty has weight 0.02, and is only for jaws and noses.

B. Comparison with DatasetGAN

We use the pre-trained StyleGAN generator on FFHQ [31] of resolution 256×256. We sample 10000 images and choose 10 images by the centers of k-means clustering on the features of the 3rd last layer of VGG [67]. The keypoints are annotated by DLIB [34], which is originally used for FFHQ alignment.

To train our model, we use the first 60000 images for training and the last 10000 images for testing. To make a fair comparison, we train our model in two different sets of few-shot examples: 1) pick from the dataset as in the main paper; 2) use the generated examples, akin to DatasetGAN.

The results are summarized in Table 4. Our model outperforms DatasetGAN in all different number of annotated examples. Note that their StyleGAN model is trained on all 70000 images in the dataset while our unlabeled dataset only contains the first 60000 images.

We remark that our model trained on the generated examples is not as good as those trained on real images. This demonstrates that the artifacts and noise in the generated images have a significant impact on the keypoint localization.

NME (%) on FFHQ dataset ↓				
Method	**Training set size**			
	1	10	20	50
DatasetGAN [99]	18.4	8.04	7.24	5.50
ours (trained with generated examples)	14.2	7.08	6.37	5.22
ours	**11.3**	**5.87**	**5.89**	**4.59**

Table 4. **Quantitative Comparison with DatasetGAN on FFHQ dataset.**

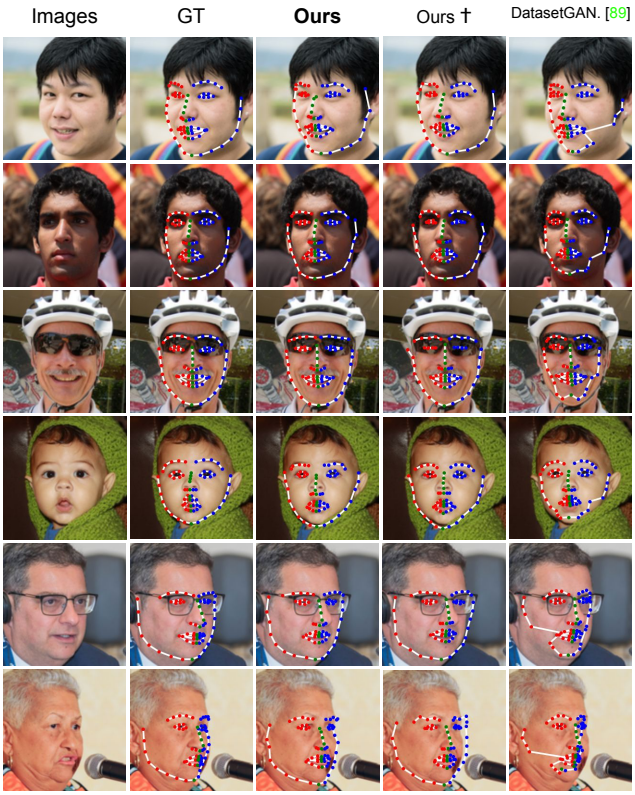

Figure 7. **Comparison with DatasetGAN**. Our model generates better shapes that are closer to the ground truth than DatasetGAN. The third row and fourth row with symbol † are obtained by training on the real annotated images and the synthetic annotated images, respectively.

C. Comparison on LSP Dataset

For completeness, we also show the results on LSP [30] dataset in Table 5. The metric and train/test split follow [51]. Our accuracy in the few-shot scenario is significantly higher than all the other baseline methods.

D. Comparison with Unsupervised Methods

In Table 6, we compare our method with state-of-the-art unsupervised methods on their commonly used datasets, i.e., 300W [63] and h36m [26]. We follow the evaluation protocol described in [28] to perform our comparisons. In particular,

PCK@0.1 (%) on LSP dataset							
Method	**Training set size**						
	1	10	20	50	5%	20%	100%
Xiao et al. [92]	**15.3**	22.1	23.9	26.5	24.0	47.8	71.1
AutoLink [20]	19.7	22.3	26.9	35.4	35.3	50.0	75.0
Moskvyak et al. [51]	3.89	7.13	15.1	27.7	**67.0**	**71.9**	74.3
ours	14.2	**49.8**	**53.2**	**58.1**	64.5	70.4	**77.9**

Table 5. **Quantitative Comparison with Baselines on LSP.**

NME (%) on H36M dataset ↓							
Method	**Training set size**						
	1	10	20	50	500	5000	100%
Xiao et al. [92]	11.8	5.53	5.33	4.71	3.44	2.35	**2.04**
Moskvyak et al. [51]	119	104	51.3	18.2	2.99	**2.30**	2.06
AutoLink (reg) [20]	18.6	10.5	9.16	7.94	3.85	3.00	2.74
AutoLink (few) [20]	9.62	5.49	4.09	3.76	3.07	2.62	2.55
Jakab et al. [28]	-	-	-	4.05	3.30	2.92	2.73
ours	**7.53**	**4.21**	**3.67**	**3.14**	**2.84**	2.57	2.58

NME (%) on 300W dataset ↓							
Method	**Training set size**						
	1	10	20	50	500	5000	100%
Xiao et al. [92]	20.2	15.0	12.8	11.0	6.37	-	**4.74**
Moskvyak et al. [51]	85.8	83.3	46.3	25.1	8.15	-	4.84
AutoLink (reg) [20]	25.4	10.7	9.90	8.31	6.07	-	5.63
AutoLink (few) [20]	15.2	9.85	9.82	8.61	6.70	-	5.25
Jakab et al. [28]	-	-	-	8.92	8.91	-	8.67
ours	**9.45**	**7.61**	**6.82**	**6.25**	**5.58**	-	4.96

Table 6. **Quantitative Comparison with Unsupervised Methods on 300W and H36M datasets.**

to match the training and evaluation scheme of unsupervised approaches [20, 28, 98] on h36m dataset, where the left and right side of the object is ambiguous during training, we flip all few-shot skeletons to facing front. At inference, we choose the correct left and right side by simply flipping the skeleton back to the correct orientation [98] if needed.

Our method outperforms the state of the art on both datasets in the [10-50]-shot scenario. For evaluating AutoLink [20], besides adding few-shot supervision as described in Section 3.2, we also follow the traditional way in unsupervised learning [20, 28, 73, 98]. Specifically, we fit a linear regression model from the unsupervised keypoints to the annotated keypoints. However, linear regression is not a data efficient approach. It requires more labels than the AutoLink (few) baseline and 100x more labels than our method, even though we chose the best L_2 regularization coefficient in [0, 10] and the optimal number of keypoints in [4,32] by grid search (10-fold cross-validation). Note that Jakab et al. [28] use unpaired annotations which are beneficial for transferring semantics across domains (e.g., sim2real) but cannot exploit to the full extent when image/pose pairs are available. What we claim as a contribution is a novel formulation that includes labeled examples and adds 3D and visibility constraints, altogether leading to substantial improvements.

PCK@0.1 (%) on WFLW dataset ↑

Method	Training set size						
	1	10	20	50	5%	20%	100%
Xiao et al. [92]	4.73	32.1	36.3	44.4	68.2	81.5	87.0
Moskvyak et al. [51]	4.25	4.85	17.7	49.6	58.1	84.9	85.4
AutoLink (few) [20]	47.7	53.4	55.0	63.1	75.5	80.2	83.7
ours	**58.3**	**71.0**	**73.9**	**77.3**	**84.1**	**86.7**	**87.5**

PCK@0.1 (%) on SynthesEyes dataset ↑

Method	Training set size						
	1	10	20	50	5%	20%	100%
Xiao et al. [92]	12.5	37.7	61.4	75.4	96.1	99.0	**99.7**
Moskvyak et al. [51]	3.09	9.74	25.2	38.8	95.0	**99.2**	99.4
AutoLink (few) [20]	25.8	49.9	73.6	82.4	95.6	98.7	**99.7**
ours	**32.9**	**77.6**	**79.4**	**89.9**	**97.7**	98.8	99.0

NME (%) on CUB-200-2011 dataset ↓

Method	Training set size						
	1	10	20	50	5%	20%	100%
Xiao et al. [92]	25.6	19.4	18.6	16.6	12.3	8.79	5.53
Moskvyak et al. [51]	64.1	55.2	51.8	39.5	12.3	7.42	**3.95**
AutoLink (few) [20]	**20.7**	18.1	16.8	14.4	10.6	8.72	5.45
ours	**20.7**	**9.94**	**9.17**	**8.97**	**6.42**	**5.19**	4.58

NME (%) on ATRW dataset ↓

Method	Training set size						
	1	10	20	50	5%	20%	100%
Xiao et al. [92]	24.1	20.1	19.5	18.3	11.7	5.61	3.09
Moskvyak et al. [51]	43.3	43.8	41.5	33.8	7.99	4.79	3.67
AutoLink (few) [20]	20.3	20.2	19.8	19.1	7.26	4.86	3.10
ours	**19.8**	**19.0**	**6.69**	**5.94**	**3.72**	**2.89**	**2.83**

NME (%) on CarFusion dataset ↓

Method	Training set size						
	1	10	20	50	5%	20%	100%
Xiao et al. [92]	29.7	22.9	21.5	18.6	15.5	8.12	4.80
Moskvyak et al. [51]	60.3	63.0	38.6	31.1	19.3	7.22	4.19
AutoLink (few) [20]	**27.0**	21.3	19.2	15.1	13.7	7.00	4.55
ours	29.0	**15.5**	**14.8**	**12.5**	**9.31**	**3.76**	**3.56**

Table 7. **Additional Quantitative Comparison with Baselines on WFLW, SynthesEyes, CUB, ATRW, and CarFusion.**

E. Additional Analysis of Jaw Landmarks

We notice that the top two jaw landmarks tend to be farther from their neighbors than the other jaw landmarks. We believe that it is due to the image reconstruction. The model tends to model the foreheads with the top two jaw landmarks so that the head structure is more clear to the model. As a result, the reconstruction has better quality.

F. More Quantitative Results

In Table 1 in Section 4, we report NME on WFLW and SynthesEyes, and PCK on CUB, ATRW, CarFusion. To promote future research, we also report additional metrics in Table 7, namely PCK on WFLW and SynthesEyes, and NME on CUB, ATRW, and CarFusion.

G. Negative Societal Impacts

Our paper has no ethical concerns but might have some potential malicious misuses on downstream tasks, such as face tracking and animation.

H. Acknowledgments

We thank Shih-Yang Su for his insightful comments on selecting core examples from a dataset, and anonymous reviewers for their valuable feedback.