SCE-MAE: Selective Correspondence Enhancement with Masked Autoencoder for Self-Supervised Landmark Estimation

Supplementary Material

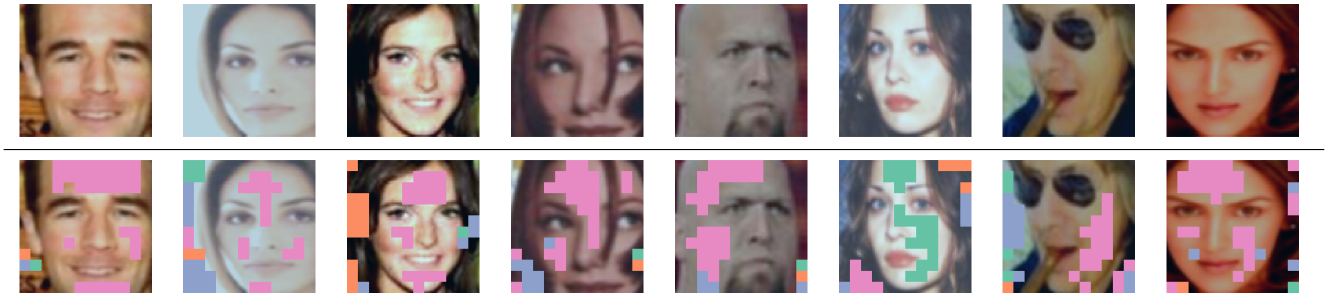

Figure S1. **Visualization of inattentive regions and clustering.** The original images are shown in the first row and the visualization results are shown in the second row. The inattentive regions are represented by colored patches and each color represents one cluster. We can see none of the landmark regions are classified as inattentive regions and semantically similar inattentive regions are grouped together.

1. Qualitative Studies on Inattentive Regions.

In Section 1 of the main paper, we highlighted the observation that the non-landmark regions (e.g., cheeks and foreheads) are larger and more uniform than the sparse and distinctive landmark regions. In Section 3.2, we explained that in order to setup selective correspondence, we first target the separation of the critical facial and the insignificant regions using the CLS token output of the MAE, and then run a simple clustering algorithm on the insignificant regions. To better understand this attentive-inattentive separation and how the clustering works, we visualize a few examples in Figure S1. The inattentive regions are denoted by colorful patches and the patches with the same color belong to the same cluster. We observe that semantically similar regions are clustered together and none of the landmark regions are classified as inattentive regions, hence corroborating our earlier hypothesis.

2. Effects of Changing Backbone Architecture

In this line of work, we are the first to adopt the Vision-Transformer [3] as the backbone architecture. In this section, we evaluate whether prior works can benefit by simply changing the backbone to ViT architectures. We switch the backbone of CL [2] and LEAD [6] and report the quantitative results on landmark matching in Table S2, and on landmark detection in Table S3. Note that both CL and LEAD rely on the extraction of *hypercolumns which require feature map of different spatial resolution.* However, the hypercolumns are not compatible with our backbones as DeiTs [7] are columnar (patch-based) architectures which can only output feature maps at the same spatial size. For this reason and for a fair comparison with our work, we evaluate the previous methods using the last layer feature from DeiTs. For landmark matching, the mean pixel error increases dramatically after changing the backbone for both the matching between same and different identities. We observe similar phenomenon on landmark detection where the performance drops on all evaluated datasets except for MAFL. Additionally, we find that the performance of CL and LEAD does not improve with a larger backbone (DeiT-S compared to DeiT-T). We attribute the performance drop to two main reasons: (1) the first stage SSL protocols of CL and LEAD, namely MoCo [5] and BYOL [4], are not designed to accommodate the requirements of vision-transformer backbone. This also explains why the performance doesn't improve after applied a larger backbone. Although integration of ViT to the MoCo framework has been addressed in MoCov3 [1], integrating MoCov3 to CL is beyond the scope of our work. (2) The use of hypercolumns is essential for CL and LEAD, however they are not available when using the DeiTs. In conclusion, naïvely switching the backbone architecture does not necessarily yield better results. The performance gain of our SCE-MAE framework over existing SOTA originates due to the intrinsic compatibility of the first-stage MAE protocol (with ViT backbones) and the ability to leverage the ViT output during the second stage.

3. Choices of Hyperparameters

3.1. Attentive Rate

We use attentive rate η to decide the portion of attentive and inattentive tokens. We study the best choice of this hyperparameter by directly dropping a certain portion of the patch tokens. The idea is that the inattentive tokens are not critical for downstream evaluation as they are mainly non-landmark

Table S2. **Quantitative evaluations on landmark matching using different backbone architectures.** We report the mean pixel error between the prediction and ground-truth on 1000 image pairs sampled from MAFL. The best and second best results are shown in **bold** and underline respectively. We group the results by backbone architecture. The error of previous SOTA methods increase dramatically when switching the backbone from (ResNet-50 + Hypercolumn) to DeiTs. This demonstrates that naïvely changing the backbone architecture does not yield better performance.

Method	Backbone	#Parameters	Same	Different
		Millions	Mean Pixel Error↓	
CL[2]	ResNet-50 + Hypercolumn	23.8	0.71	2.50
LEAD[6]	ResNet-50 + Hypercolumn	23.8	0.48	2.06
CL[2]	DeiT-T	5.4	1.31	4.32
LEAD[6]	DeiT-T	5.4	0.93	8.86
Ours	DeiT-T	5.4	0.47	1.99
CL[2]	DeiT-S	21.4	3.31	7.32
LEAD[6]	DeiT-S	21.4	0.91	8.64
Ours	DeiT-S	21.4	**0.31**	**1.69**

Table S3. **Quantitative evaluations on landmark detection using different backbone architectures.** We report the error as the percentage of inter-ocular distance on four human face datasets: MAFL, AFLW$_M$, AFLW$_R$ and 300W. For AFLW$_R$, we report the results on both the original (AFLW$_{RO}$) and corrected (AFLW$_{RC}$) datasets. We group the results by backbone architecture. We can see the performance of CL and LEAD drops when using DeiTs on all datasets except for MAFL, which demonstrates that naïvely changing the backbone architecture cannot necessarily yield better performance.

Method	Backbone	#Parameters	MAFL	AFLW$_M$	AFLW$_{RO}$	AFLW$_{RC}$	300W
		Millions	Inter-ocular Distance (%)↓				
CL[2]	ResNet-50 + Hypercolumn	23.8	2.76	6.17	5.69	5.06	4.84
LEAD[6]	ResNet-50 + Hypercolumn	23.8	2.44	6.05	5.71	5.11	4.87
CL[2]	DeiT-T	5.4	2.51	6.72	5.98	5.43	4.92
LEAD[6]	DeiT-T	5.4	2.40	6.81	6.03	5.41	5.03
Ours	DeiT-T	5.4	2.20	5.89	5.54	4.86	4.22
CL[2]	DeiT-S	21.4	2.43	6.73	5.88	5.29	4.90
LEAD[6]	DeiT-S	21.4	2.39	6.88	5.91	5.32	5.10
Ours	DeiT-S	21.4	**2.08**	**5.33**	**5.40**	**4.69**	**3.94**

regions, thus if we directly drop them, it will not affect the evaluation results much. We plot the landmark matching results at different drop rate in Figure S2. We find the elbow point at 25% to be the best choice.

3.2. Clustering

As clustering is a critical step of our proposed method, we offer some quantitative ablations in this section. There are two hyperparameters for clustering — the layer to apply clustering and the number of clusters K_c. As shown in Table S4, we first experiment with the first hyperparameter and find applying clustering after the third layer to be the best. Then we search for the best number of clusters and report the results in Table S5. We find the best choice of the cluster number to be 4.

3.3. Influence of the Correspondence Types

After attentive-inattentive separation, there are three possible correspondence types between the token pairs: attentive-attentive, attentive-inattentive and inattentive-inattentive. We study the importance of each type by setting the respective repellence hyperparameter to zero and evaluate how much the performance drops in Table S6. We find that the relationship between attentive-attentive tokens is the most important as the error increases the most when we don't enforce any repellence. This is expected as the attentive tokens covers most of the landmark regions and to distinguish between the different facial landmarks, the attentive-attentive relationship should be given more importance. We also find that the relation between attentive-inattentive to be more important than inattentive-inattentive. This is expected since the former may deliver intricate cues regarding the dependencies between the landmark and critical non-landmark regions such as landmark orientation (left vs. right) and landmark boundaries.

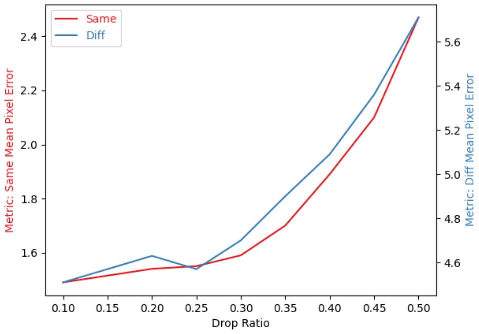

Figure S2. Landmark matching results (Mean Pixel Error) at different drop rate.

Table S4. **Landmark matching results of applying clustering after different layers.** We report the mean pixel error between the prediction and ground-truth on 1000 image pairs sampled from MAFL.

Layer	**Same**	**Diff.**	**Layer**	**Same**	**Diff.**
0	0.30	1.62	6	0.33	1.61
1	0.30	1.63	7	0.32	1.64
2	0.27	1.61	8	0.30	1.62
3	0.30	1.62	9	0.30	1.63
4	0.31	1.65	10	0.31	1.66
5	0.30	1.62	11	0.34	1.66

Table S5. **Landmark matching results of using different number of clusters.** We report the mean pixel error between the prediction and ground-truth on 1000 image pairs sampled from MAFL.

Number of clusters	**Same**	**Diff.**
1	0.33	1.68
2	0.32	1.66
4	0.27	1.61
8	0.30	1.62

Table S6. **Importance of each relationship between patch tokens.** We report the mean pixel error between the prediction and ground-truth on 1000 image pairs sampled from MAFL.

attn-attn	attn-inattn	inattn-inattn	**Same**	**Diff.**
✗	✓	✓	0.33	3.52
✓	✗	✓	0.32	1.64
✓	✓	✗	0.29	1.62
✓	✓	✓	0.27	1.61

Table S7. **Trainable components for each training stage and evaluation task.**

	Stage 1	Stage 2	Evaluation
Landmark Matching	Backbone	Projector	None
Landmark Detection			Regressor

4. More Visualizations

4.1. Visualization of Landmark Similarity Map

We visualize some of the landmark similarity maps in Figure S3. We first obtain the dense feature map from each compared method, and then computes the cosine similarity between the landmark representation and the entire feature map. We also group the results based on the property of the original image — front faces are shown in the upper rows, side faces are shown in the middle and the occluded faces are shown in the bottom. When there is occlusion or we can only see one side of the face, it is visibly difficult for the network to output discriminative representations for the occluded landmarks. As shown in Figure S3, our method generates sharper and more localized similarity map than prior arts.

4.2. Qualitative Results on Landmark Detection

Here, we show some qualitative results on landmark detection with our DeiT-B backbone in Figure S4. The model outputs accurate landmark prediction across four datasets. The model is also robust to different view angles and even some occlusions, e.g. the fifth image in MAFL.

4.3. Failure Cases of Landmark Matching

Here we visualize some failure cases of landmark matching in Figure S5. We find the main reason for these failure cases is occlusion. In some cases we can only see one side of the person's face in the image, thus the query or ground-truth landmark is occluded by other face parts. There are also cases where the landmarks are directly occluded by hand or cloth. In these cases, the query/test pixel representation at the landmark location may not effectively represent the landmark which leads to failure matching results.

5. Trainable Components for Each Stage

Our proposed method involves two training stages and the evaluation protocols for two downstream tasks are different as well. To offer a better understanding of how our framework is trained and evaluated, we detailed the trainable components for each stage and task in Table S7. Note that *only the component listed in the table is trained in the corresponding stage*, e.g., the Backbone (DeiT) is only trained in stage 1 and is frozen in all other stages.

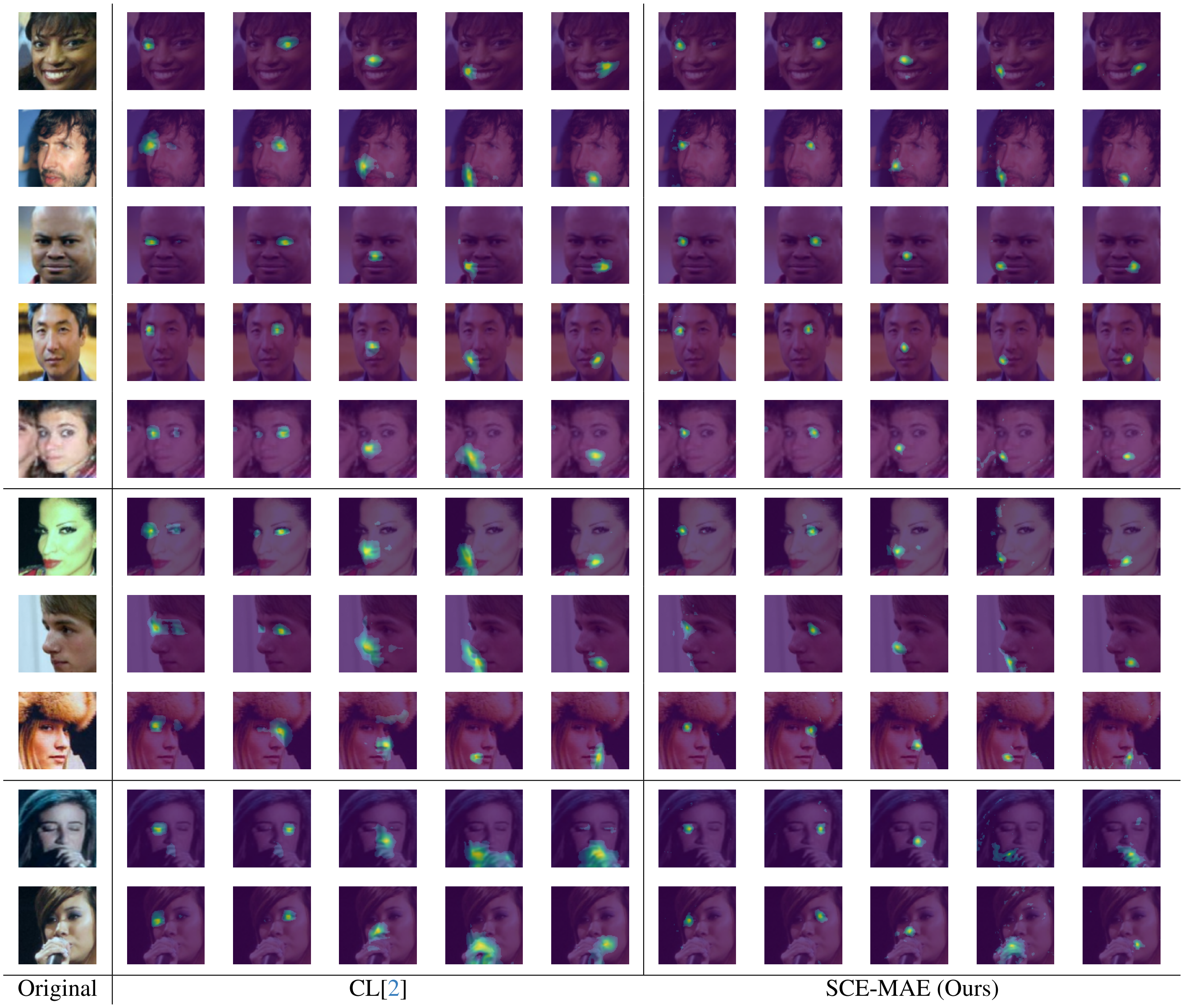

Figure S3. **Visualization of landmark similarity map.** We show the original images in the leftmost column, similarity map generated by CL in the middle and ours results on the right. We show the similarity map for each landmark and ours results are better localized at the corresponding landmark.

6. Limitations and Future Work

In this work, we presented a two-stage framework to address self-supervised face landmark estimation tasks. Despite the significant performance gain, there are still some limitations of the proposed method. Firstly, the use of the cover-and-stride technique to expand feature map resolution and produce more fine-grained representations requires additional forward passes during inference. Secondly, our second stage refining relies on the similarity map generated by the CLS token. The CLS token may be distracted when there are other salient objects in the given image. However, since our method operates on face crops, the dependency on the CLS token is mostly offloaded onto the face crop generation algorithm. Considering the limitations above, future work may involve exploring more efficient methods to gain high-resolution fine-grained feature representation and more reliable algorithms to separate the landmark and non-landmark regions. We hope our work will inspire more research in this field.

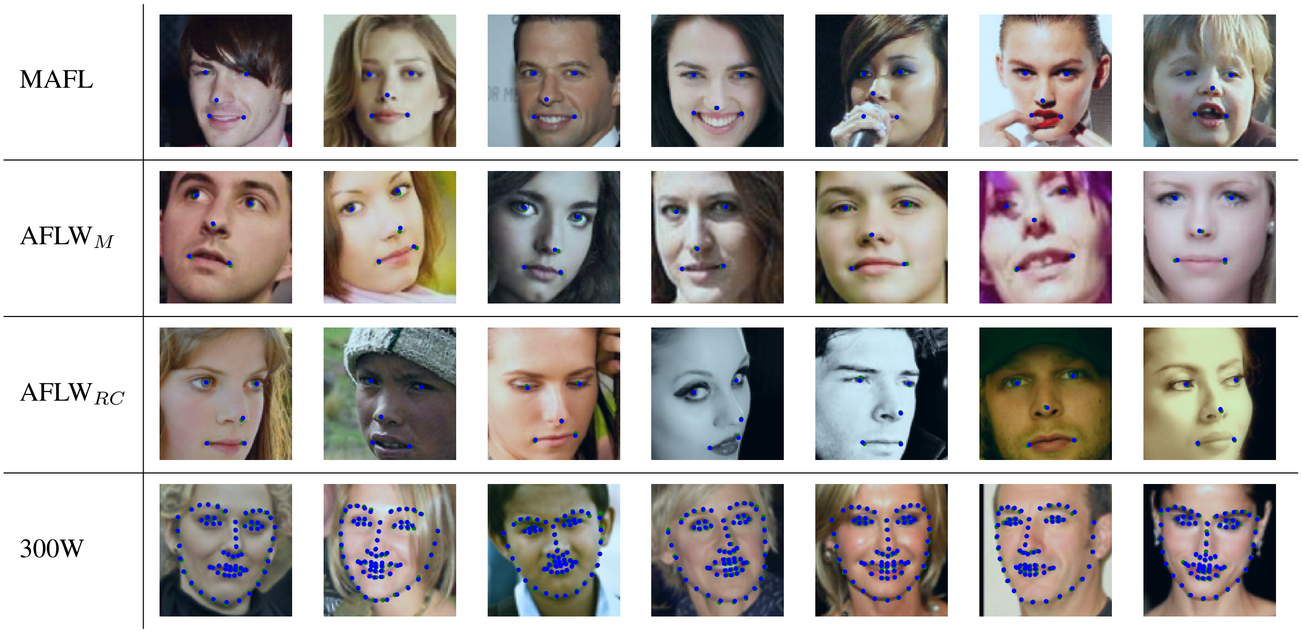

Figure S4. **Qualitative results on landmark detection.** The ground-truth and predictions are shown in green and blue dots respectively. In some cases we can only see blue dots because the prediction is almost/exact the same as ground-truth.

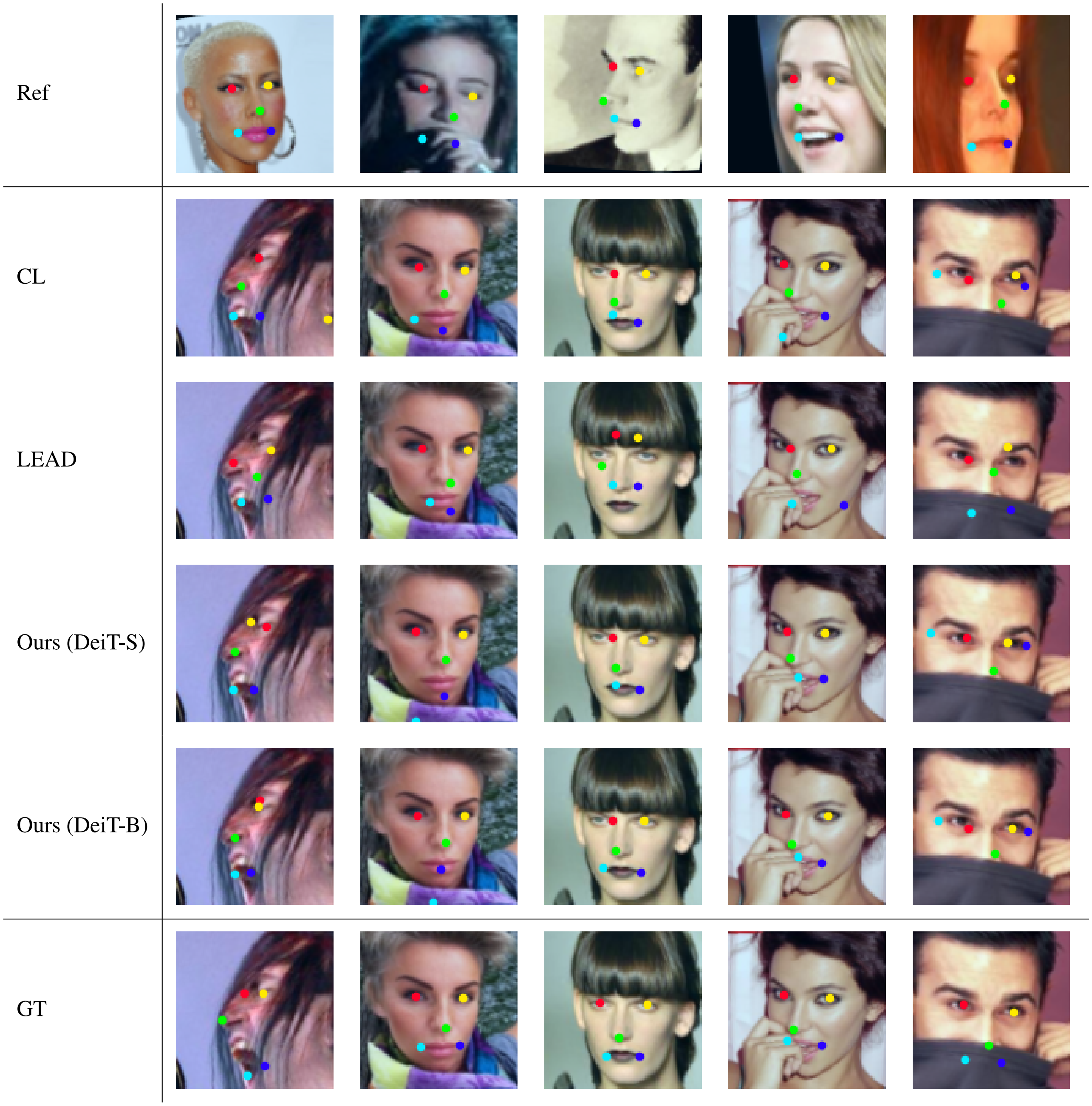

Figure S5. Failure cases of landmark matching.

References

[1] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 1

[2] Zezhou Cheng, Jong-Chyi Su, and Subhransu Maji. On equivariant and invariant learning of object landmark representations. In *ICCV*, 2021. 1, 2, 4

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[4] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, 2020. 1

[5] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1

[6] Tejan Karmali, Abhinav Atrishi, Sai Sree Harsha, Susmit Agrawal, Varun Jampani, and R Venkatesh Babu. LEAD: Self-supervised landmark estimation by aligning distributions of feature similarity. In *WACV*, 2022. 1, 2

[7] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 1