Not All Tokens Are Equal: Human-centric Visual Analysis via Token Clustering Transformer **Supplementary Material**

1. Detailed Settings for Image Classification

In this section, we provide detailed experimental settings for image classification.

We train our TCFormer on the ImageNet-1K dataset [12], which comprises 1.28 million training images and 50K validation images with 1,000 categories. We apply the data augmentations of random cropping, random flipping [14], label-smoothing [15], Mixup [20], CutMix [19], and random erasing [21]. All models are trained from scratch for 300 epochs with 8 GPUs with a batch size of 128 in each GPU. The models are optimized with the AdamW [10] optimizer, with momentum of 0.9 and weight decay of 5×10^{-2}. The initial learning rate is set to 1×10^{-3} and decreases following the cosine schedule [9]. We evaluate our model on the validation set with a center crop of 224×224 patch. The experimental settings are the same as that in [17].

2. Details of TCFormer Series

We design a series of TCFormer models with different scales for different tasks. We denote the hyper-parameters of the transformer blocks as follows and list the detailed settings of different TCFormer models in Table S1.

- R_i: The spatial reduction ratio of the transformer blocks in Stage i;
- N_i: The head number of the transformer blocks in Stage i;
- E_i: The expansion ratio of the linear layers in the transformer blocks in Stage i;
- C_i: The feature channel number of the vision tokens in Stage i.

It's worth noticing that every Clustering-based Token Merge (CTM) block contains a transformer block, whose setting is the same as the transformer blocks in the next stage.

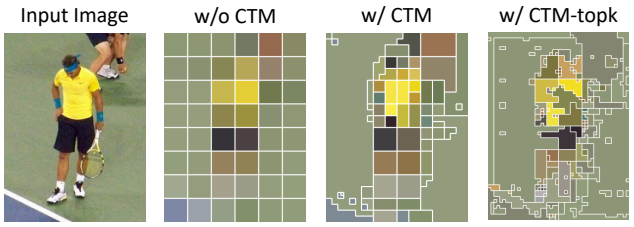

Figure S1. Example token distribution of models without CTM blocks, with CTM blocks, and with CTM-topk blocks. For the model with CTM-topk blocks, most vision tokens focus on a small part of the image area, leaving some human body parts represented by very few vision tokens or even merged with background tokens. In contrast, the vision tokens of the model with CTM blocks cover all body parts.

3. 2D Whole-body Pose Estimation

For fair comparisons with the state-of-the-art methods with larger model capacity and higher input resolution, we train TCFormer-large on the COCO-WholeBody V1.0 dataset [7] with an input resolution of 384×288. Table S2 shows the experimental results. Our TCFormer-large outperforms HRNet-w48 [13] by 1.3% AP and 1.9% AR, and achieves new state-of-the-art performance. Compared with other state-of-the-art methods, the gain of TCFormer is most obvious on the foot and hand, which are with small size in the input images. The results prove the capability of TCFormer in capturing details with small size.

4. More Ablation Studies

In this section, we show the ablation study about the clustering algorithm in the CTM block.

To validate the effect of DPC-KNN [5] algorithm, we design a variant of CTM block, which determines the cluster centers by selecting the tokens with the highest importance scores and is denoted as CTM-topk block. We build a network by replacing CTM blocks in TCFormer with CTM-topk blocks and evaluate it on the task of whole-body pose estimation.

Table S1. Detailed settings of TCFormer series. H and W denotes the height and width of input images respectively.

	Token Number	Transformer Block Setting	Block Number		
			TCFormer-Light	TCFormer	TCFormer-Large
Stage1	$\frac{H}{4} \times \frac{W}{4}$	$R_1 = 8, N_1 = 1$ $E_1 = 8, C_1 = 64$	2	3	3
Stage2	$\frac{H}{8} \times \frac{W}{8}$	$R_2 = 4, N_2 = 2$ $E_2 = 8, C_2 = 128$	1	2	7
Stage3	$\frac{H}{8} \times \frac{W}{8}$	$R_3 = 2, N_3 = 5$ $E_3 = 4, C_3 = 320$	1	5	26
Stage4	$\frac{H}{16} \times \frac{W}{16}$	$R_4 = 1, N_4 = 8$ $E_4 = 4, C_4 = 512$	1	2	2

Table S2. OKS-based Average Precision (AP) and Average Recall (AR) on the COCO-WholeBody V1.0 dataset. The baseline results are from MMPose [4]. '*' indicates multi-scale testing. ZoomNet† is trained with the COCO-WholeBody V0.5 training set.

Method	Resolution	body		foot		face		hand		whole-body	
		AP	AR	AP	AR	AP	AR	AP	AR	AP	AR
SN* [6]	480 × 480	0.427	0.583	0.099	0.369	0.649	0.697	0.408	0.580	0.327	0.456
OpenPose [1]	480 × 480	0.563	0.612	0.532	0.645	0.765	0.840	0.386	0.433	0.442	0.523
PAF* [2]	480 × 480	0.381	0.526	0.053	0.278	0.655	0.701	0.359	0.528	0.295	0.405
AE [11]+HRNet-w48 [13]	512 × 512	0.592	0.686	0.443	0.595	0.619	0.674	0.347	0.438	0.422	0.532
HigherHRNet-w48 [3]	512 × 512	0.630	0.706	0.440	0.573	0.730	0.777	0.389	0.477	0.487	0.574
ZoomNet† [7]	384 × 288	0.743	0.802	0.798	0.869	0.623	0.701	0.401	0.498	0.541	0.658
SBL-Res152 [18]	384 × 288	0.703	0.780	0.693	0.813	0.751	0.825	0.559	0.667	0.610	0.705
HRNet-w48 [13]	384 × 288	0.722	0.790	0.694	0.799	0.777	0.834	0.587	0.679	0.631	0.716
TCFormer-Large (Ours)	384 × 288	0.731	0.803	0.752	0.855	0.774	0.845	0.607	0.712	**0.644**	**0.735**

As shown in Table S3, replacing CTM blocks with CTM-topk blocks brings a significant performance drop of -7.0% AP and -7.0% AR. The performance of the model with CTM-topk blocks is even worse than the baseline without CTM blocks.

CTM-topk block determines the clustering centers based on the importance scores only, so most clustering centers are allocated to the regions with the highest scores. For the regions with middle scores, very few or even no clustering centers are allocated, which leads to information loss. As shown in Figure S1, with CTM-topk blocks, most vision tokens focus on a small part of the image area, and some body parts are represented by very few vision tokens or even merged with the background tokens, which degrades the model performance. In contrast, the clustering centers generated by the DPC-KNN algorithm cover all body parts, which is more suitable for human-centric vision tasks.

5. More Qualitative Results

In this section, we present some qualitative results for 2D human whole-body pose estimation (Figure S2), 3D human mesh reconstruction (Figure S3), and face alignment (Figure S4).

As shown in Figure S2, our TCFormer estimates the keypoints on the hand and foot accurately, which proves the capability of TCFormer in capturing the small-scale details. Our TCFormer is also capable of handling challenges including close proximity, occlusion, and pose variation. Figure S3 shows that our TCFormer estimates the human mesh accurately on the challenging outdoor images with large variations of background, illumination, and pose. As shown in Figure S4, TCFormer performs well on challenging cases with occlusion, heavy makeup, rare pose, and rare illumination.

Overall, the results show the robustness and versatility of our TCFormer.

6. More Visualizations about Token Distribution

In this section, we show the vision tokens in all stages on different tasks, *i.e.* 2D human whole-body pose estimation (Figure S5), 3D human mesh reconstruction (Figure S6), face alignment (Figure S7), and image classification (Figure S8). We observe that TCFormer progressively adapts the token distribution.

As shown in Figure S5 and Figure S6, on 2D human whole-body pose estimation and human mesh estimation

Table S3. More ablation studies on 2D human whole-body pose estimation on the COCO-WholeBody V1.0 dataset.

Method	Resolution	body		foot		face		hand		whole-body	
		AP	AR	AP	AR	AP	AR	AP	AR	AP	AR
TCFormer w/o CTM	256×192	0.667	0.749	0.562	0.695	0.617	0.621	0.479	0.590	0.535	0.639
TCFormer w/ CTM-topk	256×192	0.586	0.684	0.537	0.687	0.627	0.727	0.506	0.626	0.502	0.608
TCFormer	256×192	0.691	0.770	0.698	0.813	0.649	0.746	0.535	0.650	**0.572**	**0.678**

tasks, TCFormer merges the vision tokens of the background regions to very few tokens and pays more attention to the human body regions. For the images with simple backgrounds, such as the sky, sea, and snowfield, TCFormer merges the background tokens in stage 2 and stage 3. And for the images with complex backgrounds, distinguishing foreground from background requires high-level semantic features, so TCFormer merges the background vision tokens in the last stage. On the face alignment task, TCFormer imitates the standard grid-based token distribution in the first three stages and focuses on the face edge areas in the last stage.

We can also observe targeted token distribution on image classification. As shown in Figure S8, TCFormer allocates more tokens for the informative regions and uses fewer tokens to represent the background area with little information. In addition, the token regions generated by TCFormer are aligned with the semantic parts. This proves that TCFormer not only works on human-centric tasks but also on general vision tasks.

We also show the distribution of tokens generated with different aims. We train two models with different tasks. The first one aims to estimate only the hand keypoints, while the other one aims to estimate only the face keypoints. In Figure S9, we visualize the tokens generated by these two models, denoted as token (hand) and token (face) respectively. We find the token distribution to be task-specific, which proves that our TCFormer is able to focus on important image regions.

References

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018. 2

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2

[3] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5386–5395, 2020. 2

[4] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 2

[5] Mingjing Du, Shifei Ding, and Hongjie Jia. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems*, 99:135–145, 2016. 1

[6] Gines Hidalgo, Yaadhav Raaj, Haroon Idrees, Donglai Xiang, Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Single-network whole-body pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2

[7] 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 *Eur. Conf. Comput. Vis.*, pages 196–214, 2020. 1, 2

[8] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Brit. Mach. Vis. Conf.*, 2010. doi:10.5244/C.24.12. 5

[9] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *Int. Conf. Learn. Represent.*, 2017. 1

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *Int. Conf. Learn. Represent.*, 2019. 1

[11] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Adv. Neural Inform. Process. Syst.*, 2017. 2

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. 1

[13] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5693–5703, 2019. 1, 2

[14] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 1

[15] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 1

[16] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Eur. Conf. Comput. Vis.*, pages 601–617, 2018. 5

Figure S2. Example results of TCFormer on whole-body pose estimation.

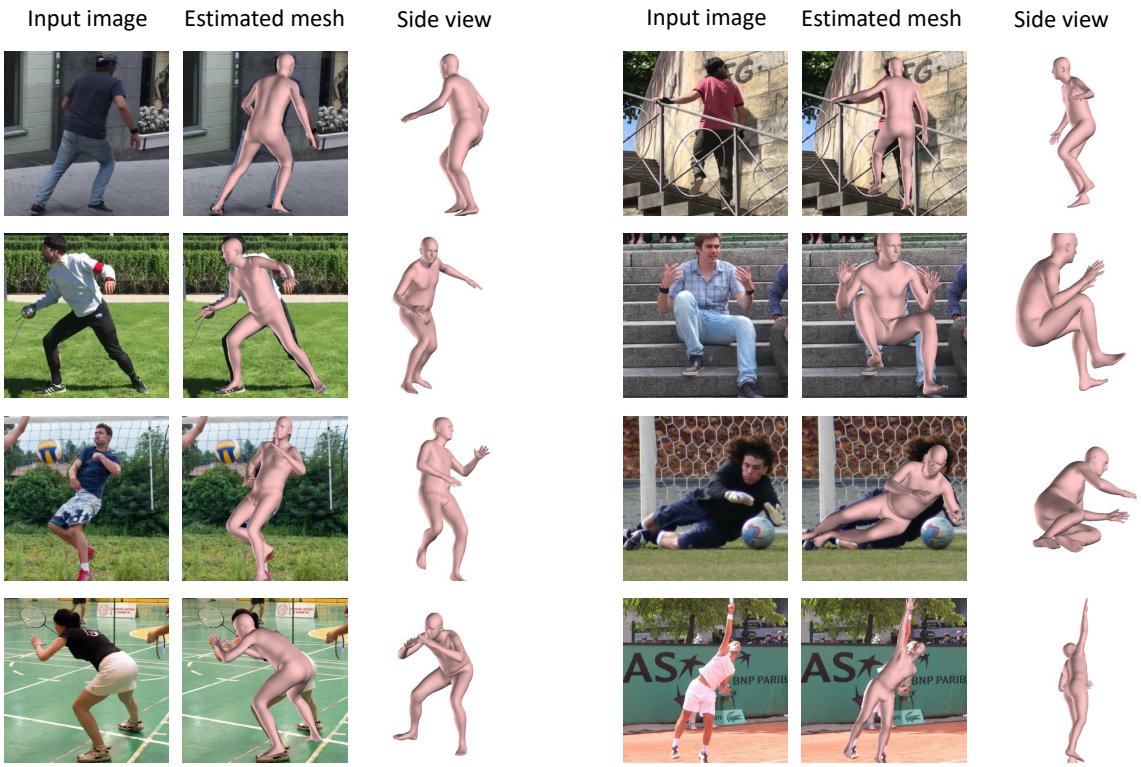

Figure S3. Example results of TCFormer on 3D mesh reconstruction. The top 2 rows are the results on the 3DPW [16] dataset and the bottom 2 rows are the results on the LSP [8] test set.

Figure S4. Example results of TCFormer on face alignment.

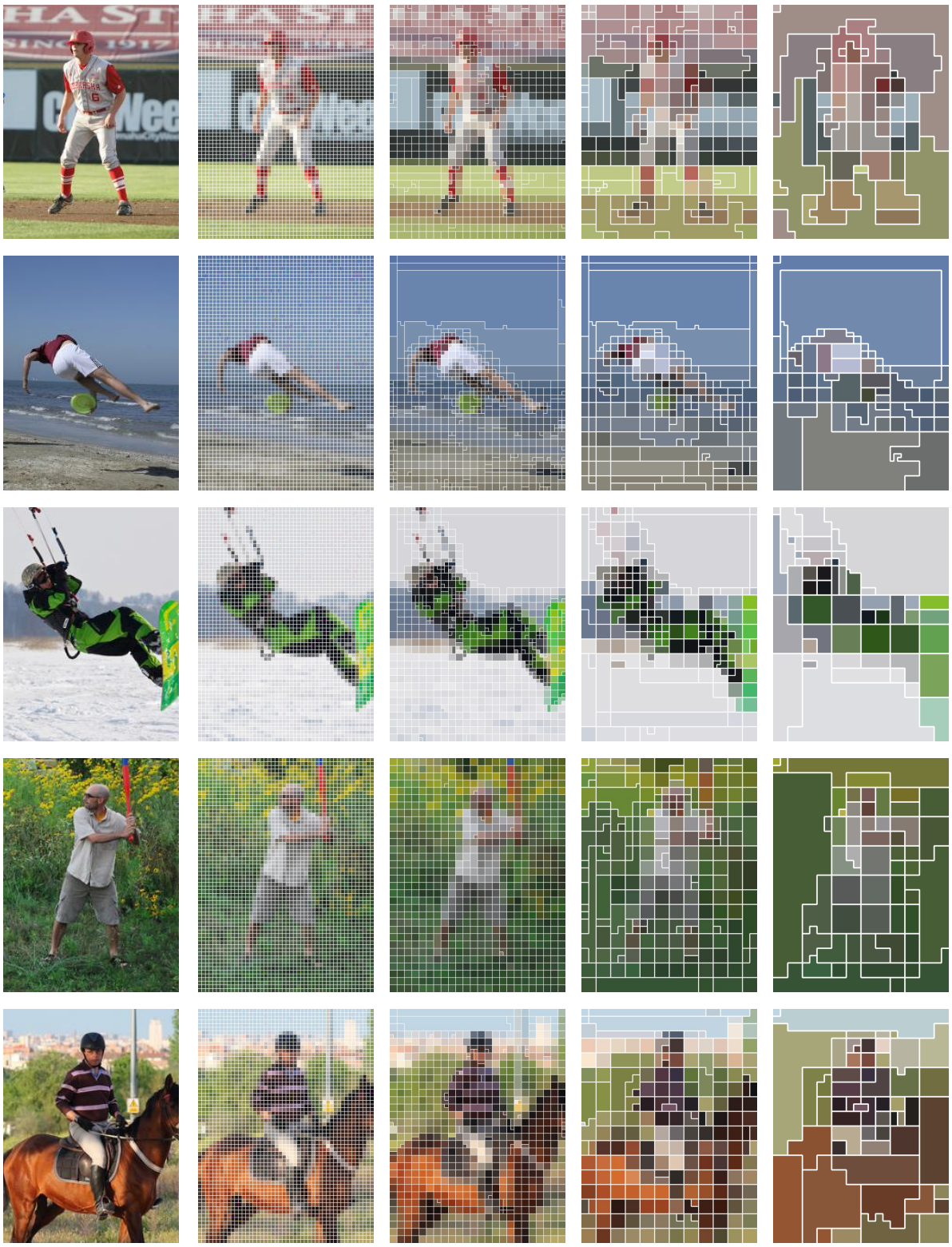

Figure S5. Example token distribution on 2D human whole-body pose estimation.

Figure S6. Example token distribution on 3D human mesh reconstruction.

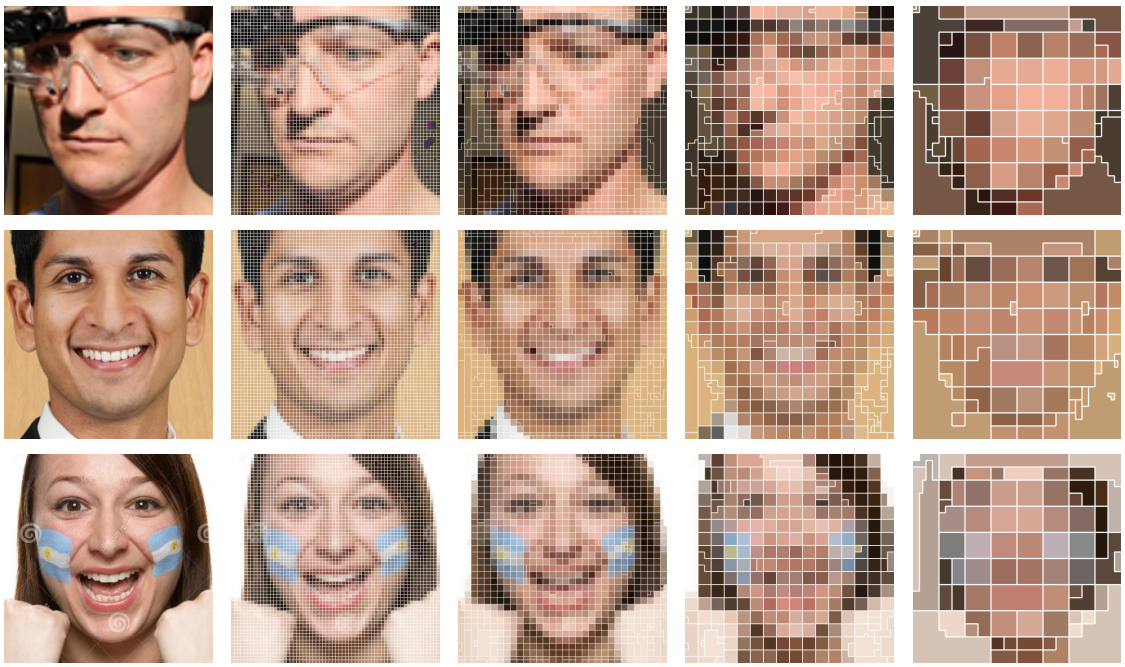

Figure S7. Example token distribution on face alignment.

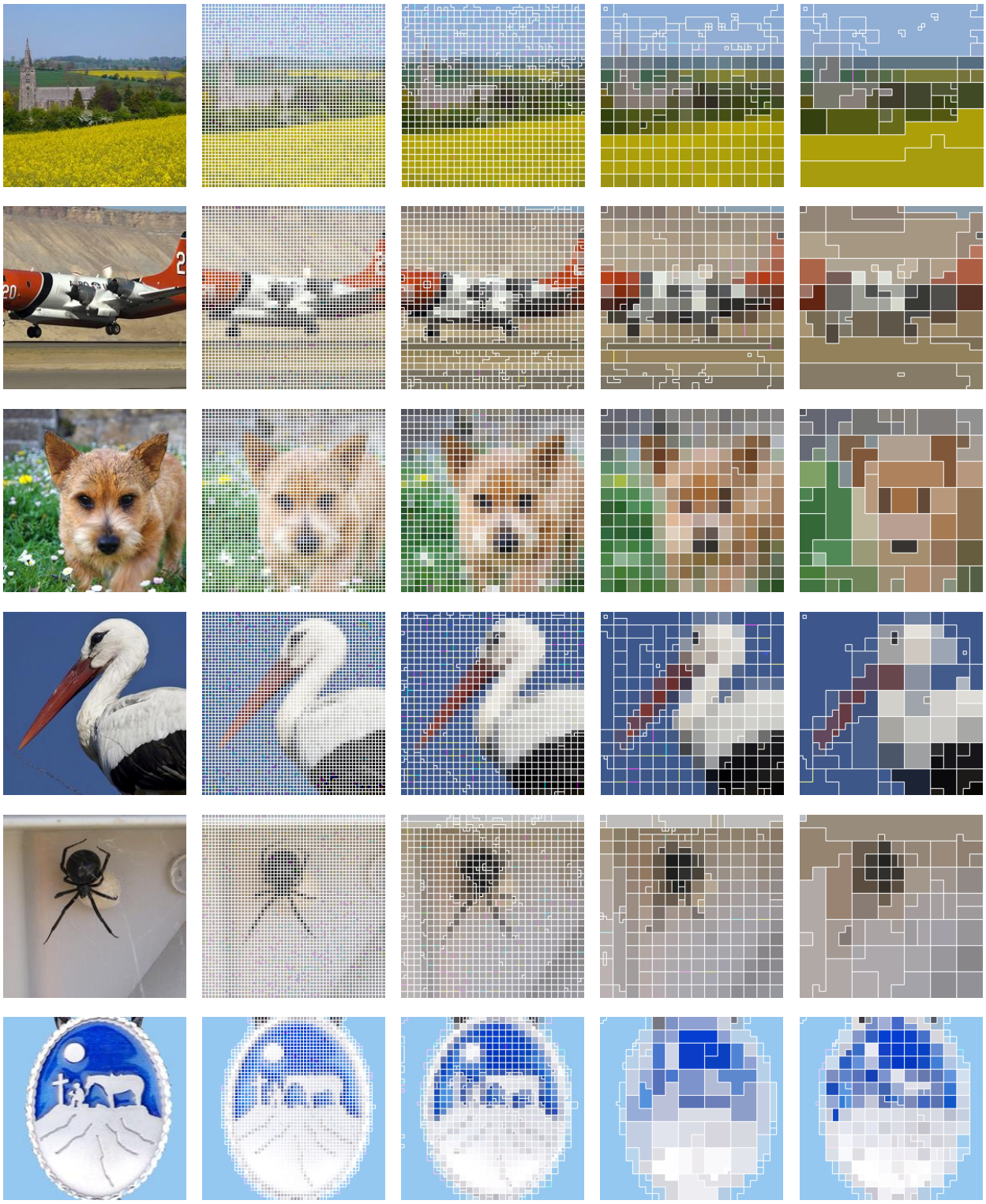

Figure S8. Example token distribution on image classification.

Figure S9. Example token distribution with different aims. We show the input image, the vision tokens generated by TCFormer that aims to estimate only hand keypoints (token (a)) and face keypoints (token (b)). TCFormer adjusts the vision token distribution according to the task.

[17] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *Int. Conf. Comput. Vis.*, 2021. 1

[18] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Eur. Conf. Comput. Vis.*, 2018. 2

[19] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Int. Conf. Comput. Vis.*, pages 6023–6032, 2019. 1

[20] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *Int. Conf. Learn. Represent.*, 2018. 1

[21] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020. 1