Supplementary Material for "Fast Adaptation for Human Pose Estimation via Meta-Optimization"

Shengxiang Hu[1], Huaijiang Sun[1]*, Bin Li[2], Dong Wei[1], Weiqing Li[1], Jianfeng Lu[1]
[1]Nanjing University of Science and Technology, Nanjing, China
[2]Tianjin AiForward Science and Technology Co., Ltd., Tianjin, China
{hushengxiang,sunhuaijiang}@njust.edu.cn

A. Network Structure

In this paper, MeTTA is implemented by a convolutional architecture for simplicity, where the backbone can replace SimpleBaseline [9] and HRNet [7] with a more complex structure. We complement the heatmap bottleneck in the primary network and the specific structure of the auxiliary network in detail.

A.1. Heatmap Bottleneck

In [3], the heatmap bottleneck is proposed to learn to extract keypoint-like structures from the input image. After the original self-supervised heatmaps $\tilde{y}^{self} \in \mathbb{R}^{W' \times H' \times K}$ are yielded, the heatmap bottleneck standardizes them into the Gaussian-like heatmaps $\hat{y}^{self}$. Specifically, each heatmap is first converted to a keypoint u_k via Softmax:

$$u_k = \frac{\sum_{u \in \Omega} u \cdot \exp(\tilde{y}_k^{self}(u))}{\sum_{u \in \Omega} \exp(\tilde{y}_k^{self}(u))}, \tag{1}$$

where we use Ω to denote the image range and u to denote the 2D coordinates in it. Then, we place the center of the Gaussian kernel with a fixed standard deviation σ on this series of keypoints $\{u_1, ..., u_K\}$ as:

$$\hat{y}_k^{self}(u) = \exp\left(-\frac{1}{2\sigma^2}\|u - u_k\|^2\right), \tag{2}$$

to obtain the self-supervised heatmaps $\hat{y}^{self}$. In MeTTA, the output of the heatmap bottleneck as pose information is used to guide our auxiliary task, *i.e.* body-specific image inpainting. During inference, minimizing the auxiliary loss can fine-tune self-supervised keypoints to match the human body in the test image.

A.2. Auxiliary Network

In MeTTA, the auxiliary network is designed to achieve body-specific image inpainting. In pursuit of generating as realistic content as possible for missing patches, most image inpainting methods have complex network structures [6] and tedious optimization processes [11]. In contrast, the main purpose of image inpainting in test-time adaptation is to provide a self-supervisory signal during inference. Thus, more important than the authenticity of reconstruction is the ability to capture human body information. In addition, the auxiliary network should be simple enough to satisfy the efficiency of fine-tuning.

Our auxiliary network consists of two encoders ϕ_{app} and ϕ_{pos}, as well as a decoder ψ, as shown in Fig. 1. Taking the masked image as input, ϕ_{app} is a fully convolutional network, which aims to extract appearance features F^{app} from the remaining pixels. The number of feature channels

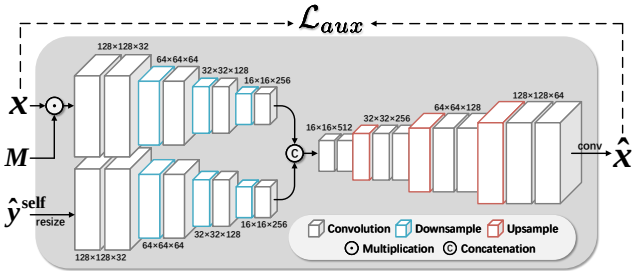

Figure 1. The structure of the auxiliary network. The first layer of the two encoders is a 7×7 convolution, and the other convolutional layers in the auxiliary network are all 3×3. Downsampling is implemented by convolution with stride of 2, and upsampling uses nearest neighbor interpolation.

Backbone	Hea.	Sho.	Elb.	Wri.	Hip	Kne.	Ank.	Mean
ResNet-101	97.3	96.1	90.9	85.9	90.0	87.3	84.2	90.6
HRNet-W32	97.4	96.2	91.5	87.7	90.5	88.6	85.4	91.4
TokenPose	97.3	96.3	91.8	88.0	90.6	88.8	85.2	91.5

Table 1. Performance of MeTTA with different backbones on MPII [1]. In general, MeTTA is able to match most heatmap-based methods, with only their last few layers modified.

*Corresponding author.

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
Penn Action [10]	86.32	86.76	87.11	87.38	87.60
Human3.6M [2]	88.50	89.24	89.74	90.05	90.16

Table 2. The accuracy after each gradient descent during test time. The total number of iterations K is set to 5 in meta-learning.

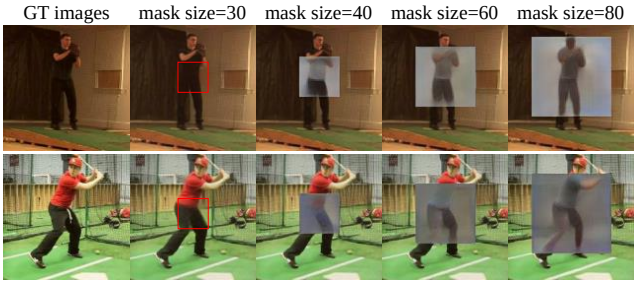

Figure 2. Results of image inpainting under restricted appearance information. After joint training, we reduce the available visual information by increasing the mask size. Thanks to the presence of pose information, there is a clear body contour in the restored image even if the human body is almost completely masked.

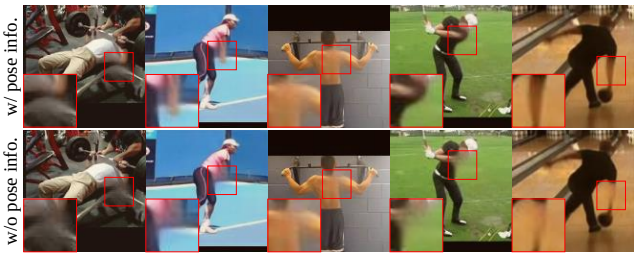

Figure 3. The effect of pose information on image inpainting. The first and second rows are the restored images with and without pose information, respectively. The results demonstrate that the pose information significantly improves the quality of restoration.

increases to 64, 128 and 256 in turn, and the resolution is at last compressed to $\frac{W}{8} \times \frac{H}{8}$. To strengthen the connection between the primary and auxiliary tasks, MeTTA uses the self-supervised heatmaps $\hat{y}^{self}$ to guide the the generation of the missing patch, which is better than simply using a shared encoder like [8]. To obtain the pose features F^{pos}, ϕ_{pos} upsamples $\hat{y}^{self}$ and then encodes it with a structure that differs from ϕ_{pos} only in the input dimension. F^{app} and F^{pos} are concatenated in the channel direction fed to ψ to predict the missing patch. The decoder ψ consists of convolution and upsampling alternately.

B. Additional Results

In this section, we conduct supplementary experiments and provide visualization to more comprehensively analyze the superiority of our MeTTA as below.

Results with different backbones. We have shown that our meta-auxiliary learning framework is compatible with convolution-based pose estimation methods, represented by SimpleBaseline [9] and HRNet [7]. Here we use TokenPose [5] as the backbone to prove that the proposed MeTTA is also suitable for Transformer-based methods, as shown in Table 1. Since the backbone used for feature extraction is frozen after joint training, the part of the computation cost involving test-time adaptation is unchanged.

Process of test-time adaptation. In our experimental setup, we demonstrate by ablation that $K = 5$ is a suitable number of updates. In order to better evaluate the process of test-time training, we give the accuracy of the meta-learned model after each gradient descent during inference on Penn Action [10] and Human3.6M [2], as shown in Table 2. The experimental results show that as the network weights are updated iteratively, the performance of the primary network is steadily improved.

Effectiveness of the auxiliary task. In our MeTTA, the self-supervised heatmaps are used as a bridge to connect the primary and auxiliary tasks. In the body-specific image inpainting task, the influence of pose information determines whether our method can accurately utilize human-related semantics to achieve test-time adaptation. To prove that the auxiliary task depends on the primary task, we debug the appearance information and pose information during image inpainting, respectively. As shown in Fig. 2, even if the appearance cues are completely removed, the model can still restore clear body contours based on the self-supervised heatmaps. Without the guidance of pose information, the restoration of the human body in the missing patch suffers from performance degradation, as shown in Fig. 3.

C. Analysis of Test-Time Adaptation in HPE

In human pose estimation (HPE), there are two works TTT [8] and TTP [4] related to test-time adaptation. In contrast, the main contribution of MeTTA is the use of meta-learning instead of multi-task learning for accurate adjustment and fast adaptation during inference. The superiority of meta-auxiliary learning has been fully explained in the main text and will not be repeated here. In addition, we propose a better auxiliary task, body-specific image inpainting. In the following, we analyze the shortcomings of the test-time adaptation methods in HPE to highlight the advantages of our auxiliary task.

In TTT [8], image rotation prediction is used to update the network weights via self-supervised learning during test time. On the one hand, only a small part of the pixels in the input image belong to the human body. On the other hand, some vision cues (such as the sky and grass) provide enough evidence for judgment. As a result, it is difficult to adjust the source model according to human-related semantics in the test image, which limits the performance of TTT.

As for TTP [4], with the help of unsupervised landmark detection [3], although remarkable performance has been made, some restrictions have been introduced. For example, TTP cannot achieve test-time adaptation based on a single image due to the need for image pairs of the same person. In addition, unsupervised landmark detection requires that two input images have the same background, which is not true in most video-based datasets. The superior performance of TTP is of little significance at the cost of the limitations of application scenarios.

D. Limitations and Discussion

MeTTA introduced in this paper uses only a single auxiliary task, body-specific image inpainting, to achieve test-time adaptation for human pose estimation. Intuitively, adopting a combination of self-supervised tasks allows for a more comprehensive perception of the differences between the source and target domains. Besides, while meta-auxiliary learning accelerates test-time adaptation, multiple iterations during inference bring a certain amount of computation. Therefore, the lightweight of the auxiliary network is one of our future work.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693, 2014. 1

[2] 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

[3] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *NeurIPS*, 2018. 1, 3

[4] Yizhuo Li, Miao Hao, Zonglin Di, Nitesh Bharadwaj Gundavarapu, and Xiaolong Wang. Test-time personalization with a transformer for human pose estimation. In *NeurIPS*, pages 2583–2597, 2021. 2, 3

[5] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *ICCV*, pages 11313–11322, 2021. 2

[6] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, pages 11461–11471, 2022. 1

[7] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 1, 2

[8] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, pages 9229–9248, 2020. 2

[9] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, pages 466–481, 2018. 1, 2

[10] Weiyu Zhang, Menglong Zhu, and Konstantinos G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, pages 2248–2255, 2013. 2

[11] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I-Chao Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. In *ICLR*, 2021. 1