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

Abstract

Domain shift is a challenge for supervised human pose estimation, where the source data and target data come from different distributions. This is why pose estimation methods generally perform worse on the test set than on the training set. Recently, test-time adaptation has proven to be an effective way to deal with domain shift in human pose estimation. Although the performance on the target domain has been improved, existing methods require a large number of weight updates for convergence, which is time-consuming and brings catastrophic forgetting. To solve these issues, we propose a meta-auxiliary learning method to achieve fast adaptation for domain shift during inference. Specifically, we take human pose estimation as the supervised primary task, and propose body-specific image inpainting as a self-supervised auxiliary task. First, we jointly train the primary and auxiliary tasks to get a pre-trained model on the source domain. Then, meta-training correlates the performance of the two tasks to learn a good weight initialization. Finally, meta-testing adapts the meta-learned model to the target data through self-supervised learning. Benefiting from the meta-learning paradigm, the proposed method enables fast adaptation to the target domain while preserving the source domain knowledge. The carefully designed auxiliary task better pays attention to human-related semantics in a single image. Extensive experiments demonstrate the effectiveness of our test-time fast adaptation.

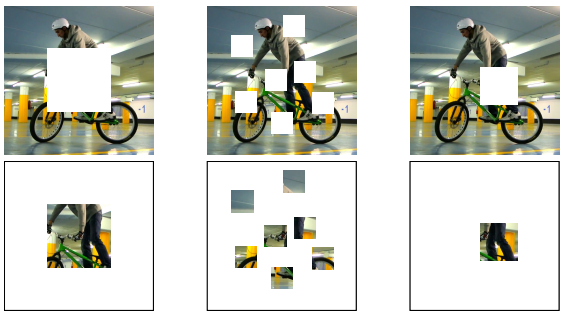
1. Introduction

Monocular 2D human pose estimation, which aims to locate body joints in images or videos, has attracted widespread attention as a fundamental task in computer vision. With the development of deep learning [30, 37, 43] and the advent of large-scale datasets [1, 11, 48], human pose estimation has made steady progress over the past few years. In supervised learning, the learned model relies heavily on the training data, which is fragile to the domain shift that is common in practical applications. Due to the limited amount of samples and the coarse division of datasets, it is inevitable that the training and test sets have differences in data distribution. Therefore, most pose estimation methods do not perform as well on the test set as they do on the training set. To solve this problem, unsupervised domain adaptation (UDA) [6, 10, 27] is proposed to improve the performance of the source model on the target domain. There are usually two requirements in practical applications, which have not been paid enough attention in the current research. Firstly, the high overhead of accessing the source domain at test time is unacceptable. Secondly, domain adaptation is expected to be completed as soon as possible. Hence, there is a pressing need for test-time fast adaptation to deal with domain shift in human pose estimation.

Test-time adaptation [34, 38, 41], *i.e*. source-free domain adaptation, adapts the source model to the unlabeled target data during inference. Recently, TTP [20] has successfully applied test-time adaptation to 2D human pose estimation to customize the model for specific instances. Concretely, TTP builds a correlation between human pose estimation and an auxiliary task (*i.e*. image rotation prediction [8] or unsupervised landmark detection [13]) by a shared encoder. Before performing keypoint localization on the test data, the network weights are updated by self-supervised learning to cope with domain shift. Since network optimization during inference depends entirely on the auxiliary task, test-time adaptation based on multi-task learning has the following drawbacks: (1) Directly tuning the pre-trained model leads to the forgetting of pose estimation knowledge. (2) The number of iterations required is large and uncertain, which is difficult for optimization. Therefore, in 2D human pose estimation, existing methods are time-consuming and their updated models are sub-optimal.

Meta-learning has become an important means to achieve source-free domain adaptation in multiple computer vision fields [3, 23, 25, 31], because meta-learning is able to learn a good weight initialization for further optimization. This allows the meta-learned model to be easily adapted to

*Corresponding author.



(a) Central region (b) Random block (c) Body-specific area

Figure 1. An example of image with different mask generation strategies. For human pose estimation, (a) central region overlooks some body joints due to spatial inductive bias. The proportion of pixels related to the human body is low, so (b) random block pays more attention to background restoration. In contrast, (c) our method captures human body information better.

a wide variety of target domains. For test-time adaptation, meta-auxiliary learning [3] guarantees that minimizing the auxiliary loss during inference is always beneficial to the primary task. In our work, we design body-specific image inpainting as the auxiliary task and apply meta-learning to couple its performance with that of human pose estimation, which overcomes the drawbacks of existing methods and achieves fast and accurate test-time adaptation.

Specifically, we first jointly train human pose estimation and the self-supervised auxiliary task to obtain a pre-trained model, because it is difficult to perform meta-learning from scratch. Then, with respect to the network weights updated by the auxiliary task, meta-training minimizes the primary loss to correlate the performance of the two tasks for a good weight initialization. Finally, meta-testing achieves fast and accurate adaptation of the source model to the target data with a definite number of iterations. As for the choice of auxiliary tasks, we propose body-specific image inpainting to assist human pose estimation, which is superior to image rotation prediction [8] and unsupervised landmark detection [13]. During the training phase, we randomly mask the area around body joints according to the ground-truth heatmaps, so that our auxiliary task focuses on human body information. During the testing phase, by sampling the predicted heatmaps, there is a tendency to mask those body joints with high confidence scores, which enhances the robustness of meta-testing. We show how our mask generation strategy differs from *central region* and *random block* in Fig. 1. It's worth noting that the proposed method is also suitable for inter-dataset domain shift. The main contributions of this paper are summarized as follows:

- We are the first to use meta-learning to achieve test-time adaptation for 2D human pose estimation. Our method enables fast and accurate adaptation to the target domain for performance gains.
- We propose body-specific image inpainting that is able to accurately capture human body information, to assist the source pose estimation model to update the network weights during inference.
- Experimental results show that our method achieves good performance due to its ability to cope with the difference in data distribution between the training and test sets.

2. Related Work

2.1. Human Pose Estimation

After continuous development, human pose estimation has become an important part of human-centered applications [42], such as action recognition [2], human parsing [46], and re-identification [29]. The performance of human pose estimation is critical for downstream tasks. Nowadays, most studies focus on the innovation of network structure to improve the accuracy of keypoint localization, among which Hourglass [30], SimpleBaseline [43], HRNet [37] and HRFormer [45] have a profound impact on human pose estimation. Due to the neglect of domain shift, existing methods perform much worse on the test set than on the training set. In light of that, we further develop human pose estimation from the perspective of domain adaptation. This is non-trivial because there are many constraints in practical applications. On the one hand, the distribution of test data is not known during the training phase. On the other hand, timeliness requires domain adaptation to be done as soon as possible. Given these conditions, test-time adaptation is an ideal scheme to deal with the difference in data distribution between the training and test sets.

2.2. Test-Time Adaptation

As a type of unsupervised domain adaptation, test-time adaptation accesses only unlabeled test data to fine-tune the source model during inference. Due to the nature of source-free, test-time adaptation has attracted attention in computer vision fields, including image classification [40, 41], image dehazing [23, 44], and semantic segmentation [18, 26]. In contrast, human pose estimation is more complex than the above tasks, so test-time adaptation for it has not been fully studied. Existing methods are mainly through self-training or self-supervision to update the network weights. TTP [20] correlates supervised and self-supervised pose estimation in the form of multi-task learning. During inference, the shared encoder is updated by the self-supervised branch. In [34], the source-protect module distills source knowledge to constrain the optimization direction, and the target-relevant module adapts to the target domain via exponential moving average. POST [36] ensures the anatomical rationality of pseudo-labels based on a learned pose prior. Consistency regularization with respect to image transformations is used as a criterion for test-time adaptation. However, the number

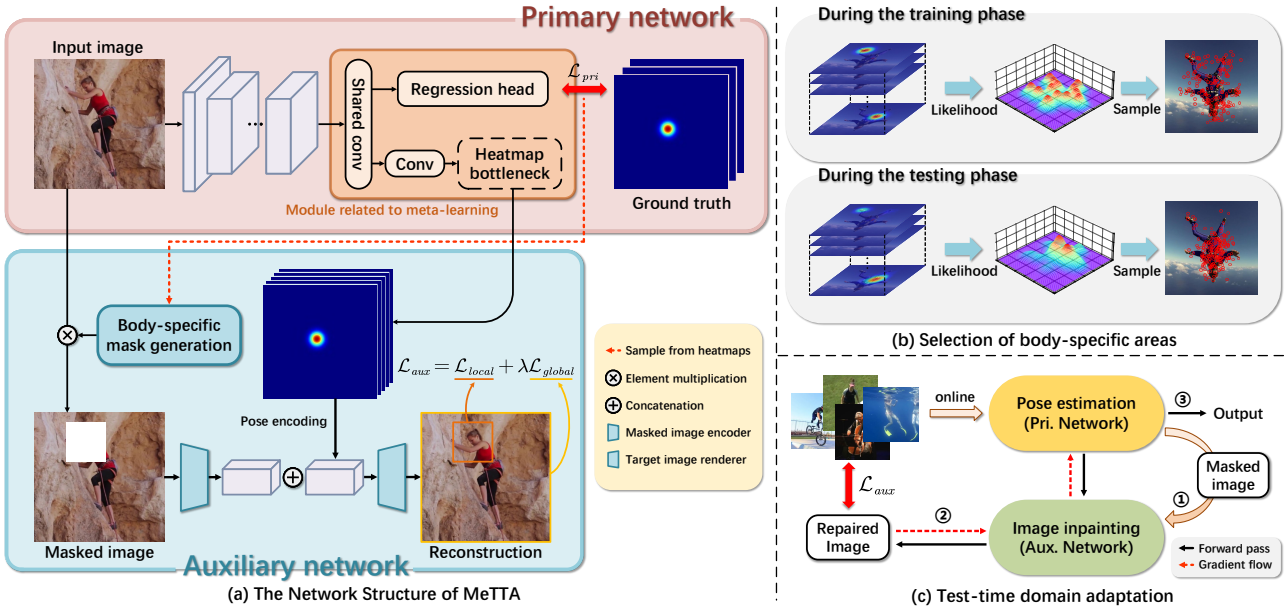


Figure 2. Overview of our test-time fast adaptation method. MeTTA takes human pose estimation as the primary task and body-specific image inpainting as the auxiliary task. During the training phase, the center of the region mask is sampled from the ground-truth heatmaps. To cope with domain shift, test-time training is performed in three steps: First, the meta-learned weights are used to output the predicted heatmaps to determine the location of the region mask. Then, the primary network is updated by self-supervised learning from the auxiliary task. Finally, the source model after fast adaptation achieves more accurate localization than the pre-trained model in the target domain.

of iterative steps required to converge in existing methods is numerous and uncertain, which limits performance due to catastrophic forgetting and does not meet the requirements for high efficiency. To accelerate test-time adaptation for human pose estimation, we propose a novel meta-auxiliary learning method. Compared to the pre-trained model in previous works [20, 38], our meta-learned model has the ability to complete fine-tuning with few iterations.

2.3. Meta-Auxiliary Learning

Meta-learning, also known as learning to learn, is a broad concept used to acquire learning algorithms for specific meta-knowledge. Lately, weight initialization represented by model-agnostic meta-learning (MAML) [5] has been widely used [4, 9, 35], since a good weight initialization can avoid local optimal and accelerate convergence. Due to these advantages, meta-learning is associated with auxiliary learning to achieve test-time adaptation, which has shown remarkable results on multiple visual tasks. For dynamic scene deblurring, [3] adopts image reconstruction to quickly capture and adapt to internal information in test images. In [23], the helper network is used to learn the haze patterns in the source domain and update the dehazing network with few iterations during inference. Similarly, [24] uses image reconstruction as self-supervision to achieve fast adaptation for future depth prediction. In 2D human pose estimation, compared to the previous work [20], we use meta-auxiliary learning instead of multi-task learning to achieve test-time adaptation, in which a better auxiliary task, body-specific image inpainting, is proposed. Given the test image, after a small and definite number of iterations are performed by self-supervised learning, our pose estimation model obtains a significant performance improvement.

3. Proposed Method

In this section, we describe the proposed method to deal with domain shift, called MeTTA, in which body-specific image inpainting is used to assist human pose estimation to achieve test-time adaptation. This well-designed auxiliary task is able to accurately capture human body information from a single image. Compared to the previous test-time adaptation methods [20, 34, 36], introducing meta-auxiliary learning brings two benefits: accurate adjustment and fast adaptation. An overview of MeTTA is shown in Fig. 2.

3.1. Network Architecture

The proposed MeTTA contains a primary network used for human pose estimation and an auxiliary network for image inpainting, as shown in Fig. 2(a). Our method provides a simple and effective framework for fast adaptation during test time, which is compatible with most pose estimation methods. We introduce the purpose and implementation of these two networks as follows:

The primary network. Given an input image x of size $W \times H \times 3$, the mainstream pose estimation methods predict

the heatmaps of all body joints $\hat{y} \in \mathbb{R}^{H' \times W' \times J}$ for keypoint localization, where $W', H' = \frac{W}{4}, \frac{H}{4}$. The 2D coordinates are calculated through a post-processing operation.

In the primary network, we use SimpleBaseline [43] and HRNet [37] as the backbone to extract feature maps. We attach a convolution-based Y-shaped structure used to yield supervised and self-supervised heatmaps. After the shared convolution, a branch acts as a regression head to output the predicted heatmaps $\hat{y}$. The primary loss is formulated as:

$$\mathcal{L}_{pri} = \|\hat{y} - y\|^2, \tag{1}$$

where y is the ground-truth heatmaps. If the rest of MeTTA is ignored, it is no different from general pose estimation methods. In the other branch, the self-supervised heatmaps $\hat{y}^{self} \in \mathbb{R}^{W' \times H' \times K}$ are obtained by a 1×1 convolution and a heatmap bottleneck, where K denotes the number of keypoints. For domain adaptation, [12, 40] demonstrate that significant performance gains can be achieved by adjusting just a few layers in the network. Since it is time-consuming to update the entire primary network during inference, only the Y-shaped structure is involved in meta-optimization and the backbone is fixed after joint training (see Sec. 3.2 for details). For the sake of description, we uniformly denote the learnable weights in the primary network as θ_{pri} without further distinction.

The auxiliary network. As a common pretext task to learn visual representations, image inpainting [33] aims to restore missing patches in masked images. Since no additional manual annotation is required, image inpainting is chosen as our auxiliary task to update the network weights according to the test data. Nowadays, *central region* and *random block* are the two main mask generation strategies to remove one or more regions. However, these region masks are determined independently of the input image, which makes it difficult for image inpainting to focus on human body information. To better assist 2D human pose estimation, we design a body-specific image inpainting task as shown in Fig. 1. Specifically, given an input image and its heatmaps, we randomly sample the likelihood that each pixel has human body information to determine the location of the region to be removed. We calculate the likelihood as:

$$p(u, v) = \frac{\max_j h(u, v, j)}{\sum_{v=1}^{H'} \sum_{u=1}^{W'} \max_j h(u, v, j)}, \tag{2}$$

where $p(u, v)$ denotes the probability of sampling at (u, v). For the training data, we use its ground-truth heatmaps y as $h(u, v, j)$, where j indicates the j-th keypoint. This means that every visible body joint is considered equally in the selection of the mask region, which is conducive to learning comprehensive human body information. For the test data, we take the predicted heatmaps $\hat{y}$ as $h(u, v, j)$ to calculate the likelihood. As a result, the mask regions are more likely to be located near those body joints with high confidence, which makes the auxiliary task more reliable to improve human pose estimation during inference. The distribution of the masked regions is illustrated in Fig. 2(b). Compared to the auxiliary tasks [8, 13] used in TTP [20], body-specific image inpainting focuses better on human body information based on a single image.

Taking the masked image as input, the auxiliary network f_{aux} performs image inpainting under the guidance of a pose prior. Concretely, there are two encoders ϕ_{app} and ϕ_{pos} to obtain appearance features F^{app} from the masked image and pose features F^{pos} from the self-supervised heatmaps $\tilde{y}$. Then F^{app} and F^{pos} are concatenated and fed into a decoder to predict the missing patch. The auxiliary loss consists of a local term and a global term, where the region mask is denoted as M:

$$\mathcal{L}_{aux} = \|M \odot (\hat{x} - x)\|_2^2 + \lambda \text{PerceptualLoss}(\hat{x}, x) \tag{3}$$

where $\hat{x} = f_{aux}((1 - M) \odot x, \hat{y}^{self})$. The MSE loss for the missing patch prompts the learning of human-relevant semantics and the Perceptual loss [15] for the entire image enforces the global consistency of reconstruction.

3.2. Test-Time Fast Adaptation

In human pose estimation, slow convergence is a common shortcoming of existing test-time adaptation methods [20, 34, 36]. To solve this problem, meta-auxiliary learning is proposed to achieve fast adaptation. In MeTTA, each image is treated as a single "task" in which we force the network weights updated by the auxiliary network to minimize the primary loss. In this way, the number of iterations for fine-tuning is a predefined parameter, which makes our test-time adaptation deterministic and controllable. Compared to the baseline TTP [20], our meta-learned model from the source domain, as a better initialization than pre-trained models, adapts more easily to the test data (*i.e.* new "task"). What's more, meta-auxiliary learning correlates the performance of the primary and auxiliary tasks. This ensures that human pose estimation during inference always benefits from self-supervised learning. The optimization process consists of three stages: joint training, meta-training, and meta-testing, and we introduce them below.

Joint training. It is quite difficult to train deep neural networks from scratch using meta-learning. Instead, in the beginning, we jointly train the primary and auxiliary tasks on the source domain. After that, a pre-trained model is obtained to prepare for subsequent meta-optimization. The loss function for joint training is formulated as:

$$\mathcal{L} = \mathcal{L}_{pri}(\hat{y}, y; \theta_{pri}) + \mu \mathcal{L}_{aux}(\hat{x}, x, M; \theta_{pri}, \theta_{aux}), \tag{4}$$

where hyper-parameter $\mu \in (0, 1]$ is the weight coefficient used to balance the primary and auxiliary tasks.

Meta-training. In order to accelerate adjustment and improve performance for test-time adaptation, it is crucial to find a good weight initialization that is easy to approach the global optimal of various domains. For the training data (x_i, y_i, M_i), the process of updating the network weights by self-supervised learning is given by:

$$\hat{\theta}_i \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{aux}(\hat{x}_i, x_i, M_i), \tag{5}$$

where α denotes the adaptation learning rate in the inner loop. For illustration, $\hat{\theta}_i$ is the network weights adjusted once by gradient descent. Actually, the above process is iteratively executed several times.

So far, there is no difference between our method and multi-task learning for test-time adaptation. Without extra constraints, the direction of optimization during inference is entirely determined by the auxiliary task rather than human pose estimation. In this case, no matter how well-designed auxiliary tasks are, the difference between the primary and auxiliary tasks inevitably leads to error accumulation over time. To ensure that $\mathcal{L}_{aux}$ correctly guides the update of network weights to improve the performance of keypoint localization, the meta-objective is defined as:

$$\arg\min_{\theta} \sum_{i=1}^{N} \mathcal{L}_{pri}(\hat{y}_i, y_i; \hat{\theta}_i), \tag{6}$$

where N represents the number of batch size. It is worth noting that the primary loss $\mathcal{L}_{pri}$ is calculated for $\hat{\theta}_i$, while the optimization is performed on θ. The meta-objective is achieved by gradient descent as follows:

$$\theta \leftarrow \theta - \beta \sum_{n=1}^{N} \nabla_\theta \mathcal{L}_{pri}(\hat{y}_i, y_i; \hat{\theta}_i), \tag{7}$$

where β denotes the meta-learning rate in the outer loop. The overall procedure of meta-training is summarized in Algorithm 1. Since updating the entire model incurs a large computational cost, we freeze the backbone of the primary network during meta-learning.

Meta-testing. Starting with the meta-learned weights, the source model easily adapts to the target data with few iterations. In the meta-testing stage, we randomly mask the test image based on its predicted heatmaps as input to our MeTTA. The auxiliary network adjusts θ based on the reconstruction results of the missing patch. The updated model is used to perform human pose estimation on the test image, as shown in Fig. 2(c). Due to the constraint of the outer loop in meta-optimization, the design requirements for auxiliary tasks (*i.e.* relevance to the primary task) are relaxed to some extent, which makes MeTTA both simple in structure and superior in performance.

Algorithm 1: Meta-training

Require: Pre-trained networks f_{pri} and f_{aux}
Require: Learning rates α and β
Output: Meta-learned network weights θ

while *not converged* **do**
  Sample a batch of training data $\{x_i, y_i\}_{i=1}^{N}$
  for *each* x_i **do**
    Generate the body-specific mask M_i
    Update the network weights:
    $\hat{x}_i = f_{aux}((1-M_i) \odot x_i, \hat{y}_i^{self} = f_{pri}(x_i))$
    $\hat{\theta}_i \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{aux}(\hat{x}_i, x_i, M_i)$
  end
  Minimize $\mathcal{L}_{pri}$ by gradient descent:
  $\theta \leftarrow \theta - \beta \sum_{n=1}^{N} \nabla_\theta \mathcal{L}_{pri}(\hat{y}_i = f_{pri}(x_i), y_i; \hat{\theta}_i)$
end

4. Experiments

4.1. Datasets and Evaluation Metrics

Our experiments on the Penn Action [48], Human3.6M [11], and MPII [1] datasets show that MeTTA narrows the performance gap of pose estimation models between the training and test sets. Unlike most existing methods that focus on innovation in network structure, taking the impact of data distribution into account, we improve human pose estimation from the aspect of optimization. As a test-time adaptation method, MeTTA is also suitable for inter-dataset domain shift (*e.g.* SURREAL [39] → Human3.6M [11] and SURREAL [39] → LSP [16]). We adopt the Percentage of Correct Keypoint (PCK) metric to evaluate the accuracy of keypoint localization.

Penn Action [48] is a single-person video dataset that consists of 2326 sequences and involves 15 activities. The human body is represented as 13 keypoints annotated with 2D locations. In Penn Action, there are 1258 videos as the training set and 1068 videos as the test set.

Human3.6M [11] is a standard dataset for human pose estimation and contains 3.6 million images with 2D pose annotations for 17 pose joints. We follow the standard protocol to take 5 subjects (S1, S5, S6, S7, S8) as the training set and 2 subjects (S9, S11) as test set.

MPII [1] is a large-scale image-level dataset containing around 25k images and over 40k person instances. These images cover 410 human activities with pose annotations for 16 body joints. Following the standard split, we take 22k samples for training and 3k samples for testing.

SURREAL [39] is a synthetic human pose dataset that is used as a source dataset for domain adaptation in previous works [14, 34, 36]. The photo-realistic videos are generated from human motion capture data with indoor scenes as the background. There are over 6 million frames in it.

Method	Venue	Backbone	Hea.	Sho.	Elb.	Wri.	Hip	Kne.	Ank.	Mean
Conventional pose estimation methods										
SimpleBaseline [43]	ECCV'18	ResNet-101	96.9	95.9	89.5	84.4	88.4	84.5	80.7	89.1
HRNet [37]	CVPR'19	HRNet-W32	97.1	95.9	90.3	86.4	89.1	87.1	83.3	90.3
DARK [47]	CVPR'20	HRNet-W32	97.2	95.9	91.2	86.7	89.7	86.7	84.0	90.6
PRTR [19]	CVPR'21	HRNet-W32	97.3	96.0	90.6	84.5	89.7	85.5	79.0	89.5
TokenPose [21]	ICCV'21	HRNet-W48	97.1	95.9	90.4	86.0	89.3	87.1	82.5	90.2
SimCC [22]	ECCV'22	HRNet-W32	97.2	96.0	90.4	85.6	89.5	85.8	81.8	90.0
Posur [28]	ECCV'22	HRNet-W32	–	–	–	–	–	–	–	90.5
Test-time adaptation methods										
TTT [38]	ICML'20	ResNet-101	96.9	96.0	89.8	84.8	88.9	84.9	81.0	89.4
MeTTA(Ours)	–	ResNet-101	97.3	96.1	90.9	85.9	90.0	87.3	84.2	90.6
MeTTA(Ours)	–	HRNet-W32	97.4	96.2	91.5	87.7	90.5	88.6	85.4	91.4

Table 1. Evaluation results on MPII [1]. By test-time adaptation to deal with domain shift, we bring SimpleBaseline and HRNet to the level of state-of-the-art conventional pose estimation methods. Given a test image, MeTTA achieves better performance than TTT due to our more suitable auxiliary task and the introduction of meta-learning. We indicate the best value in red and the second best value in blue.

Method		TA	MO	Penn Action	Human3.6M
SimpleBaseline [43]				85.23	85.42
video-based	TTP [20]	✓		87.27(+2.04)	89.43(+4.01)
	TTP† [20]	✓		87.75(+2.52)	91.70(+6.28)
image-based	TTT [38]	✓		85.86(+0.63)	88.01(+2.59)
	MeTTA	✓		86.67(+1.44)	89.50(+4.08)
	MeTTA	✓	✓	87.60(+2.37)	90.16(+4.74)

Table 2. Comparison of MeTTA and existing test-time adaptation methods on Penn Action [48] and Human3.6M [11]. TA: test-time adaptation. MO: meta-optimization. † indicates that TTP adopts Transformer design. To be fair, all models take ResNet-50 as the backbone and do not use flip test during adaptation.

LSP [16], full name Leeds Sports Pose, is a real-world human pose dataset. There are 2k images collected from the wild, covering various sports activities. In our work, we use the LSP dataset as the target domain that is clearly different from SURREAL in terms of data distribution.

4.2. Implementation Details

We set the batch size to 32 and use the Adam optimizer [17] for joint training. The hyper-parameters λ and μ are set to 0.01 and 0.001. In experiments with intra-dataset domain shift, the base learning rate is initialized to 1×10^{-3} and decays twice at a rate of 0.1 in joint training. We use the learning schedule [90, 120, 140] in epochs for Penn Action and MPII, and [45, 60, 70] in epochs for Human3.6M. In meta-learning, we fix the learning rate $\alpha = \beta = 1 \times 10^{-4}$, and perform 5 iterations through SGD for fast adaptation. In experiments with inter-dataset domain shift, following the previous work [36], we set the base learning rate to 1×10^{-4} and train our MeTTA for 30 epochs on SURREAL, with 500 iterations executed in per epoch. The base learning rate decays to 1×10^{-5} after 5 epochs and to 1×10^{-6} after 20 epochs. In meta-learning, the learning rate α and β are set to 1×10^{-5}. After 5 iterations, the meta-learned source model fast adapts to the test data from LSP and Human3.6M. We set the size of the region mask to 30 × 30 for the 128 × 128 input image and to 60 × 60 for the 256 × 256 input image. Our experiments are conducted on an open-source machine learning, PyTorch [32].

4.3. Results under Intra-Dataset Domain Shift

We compare the proposed MeTTA with the conventional pose estimation methods on MPII [1], as shown in Table 1. Note the fact that even in the same dataset, the training data and the test data are not identical in distribution. From this perspective, our method uses test-time adaptation to handle domain shift for performance gains, which enables two simple pose estimation models SimpleBaseline [43] and HRNet [37] to achieve superior performance. In recent studies of human pose estimation, Transformer has been sought after due to its long-range modeling capabilities. However, at the cost of excellent performance, Transformer-based models require a long training process and a large amount of data, which is not friendly for most practical applications. The experimental results show that for human pose estimation, the improvement brought by domain adaptation is no less than that brought by structural innovation.

Further, we compare MeTTA with existing test-time adaptation methods on Penn Action [48] and Human3.6M [11], as shown in Table 2. Although TTP [20] makes the source model perform well on the target data, there is a harsh constraint: unsupervised landmark detection [13] as the auxiliary task limits the input to video with the same background. TTT [38] implements test-time adaptation for a single image, but it cannot guarantee that self-supervised learning from image rotation prediction [8] relies on human semantics rather than other visual cues. In our MeTTA,

Method	OL	Sld.	Elb.	Wri.	Hip	Kne.	Ank.	All
Source-only	–	51.5	65.0	62.9	68.0	68.7	67.4	63.9
MMT [7]	×	60.9	70.9	70.3	81.1	79.3	72.8	71.5
RegDA-SF [14]	×	54.8	70.5	67.6	65.4	73.2	70.0	66.5
POST [36]	×	66.5	83.9	81.0	84.6	83.1	82.6	80.3
SP+TR [34]	×	70.7	85.4	83.8	86.6	85.2	85.0	83.2
MeTTA	✓	71.0	84.7	83.2	85.4	84.4	84.1	82.5

Table 3. PCK@0.05 on SURREAL [39] → LSP [16]. OL: online.

Method	OL	Sld.	Elb.	Wri.	Hip	Kne.	Ank.	All
Source-only	–	69.4	75.4	66.4	37.9	77.3	77.7	67.3
MMT [7]	×	73.2	83.5	72.4	45.1	80.8	83.9	73.9
RegDA-SF [14]	×	70.6	82.0	69.8	43.3	79.1	79.4	71.5
POST [36]	×	81.3	88.5	77.4	46.1	83.4	83.4	76.7
SP+TR [34]	×	77.9	88.8	80.4	52.3	84.2	86.9	78.7
MeTTA	✓	78.1	88.6	80.6	51.3	83.8	86.5	78.5

Table 4. PCK@0.05 on SURREAL [39] → Human3.6M [11].

body-specific image inpainting is used to assist human pose estimation. In contrast, this auxiliary task focuses more on human body information in a single image. To ensure that self-supervised learning is always beneficial for human pose estimation during inference, we deeply couple the primary and auxiliary tasks via meta-optimization. For intra-dataset domain shift, MeTTA is superior to existing image-based methods and comparable to video-based TTP.

4.4. Results under Inter-Dataset Domain Shift

Although the original intention of this paper is to address performance degradation caused by domain shift within a dataset, our method is also valid for inter-dataset domain shift. Recently, deploying pose estimation models trained on synthetic datasets to real-world data through domain adaptation, which has been an effective strategy to reduce the need for manual annotation. In human pose estimation, four SOTA source-free domain adaptation methods MMT [7], RegDA-SF [14], POST [36] and SP+TR [34] are used for comparison with our method. The experimental results from SURREAL [39] to LSP [16] and to Human3.6M [11] are given in Table 3 and Table 4. In the above methods, domain adaptation is designed for the entire target domain, which requires access to all test data or even multiple times until domain adaptation is completed. In contrast, MeTTA obtains comparable performance while also enabling online adaptation to one or a batch of images.

4.5. Ablation Study

In this section, we investigate the number of inner loops in meta-optimization, which dictates how many updates needed in test-time fast adaptation. For the auxiliary task in MeTTA, we implement ablation experiments to study the impact of mask size and the design of auxiliary loss.

Method	MO	Penn Action	Human3.6M	GFLOPs
$K = 0$	–	85.74	87.49	2.15
$K = 1$		86.18	88.31	16.59
	✓	86.48	88.66	
$K = 5$		87.12	89.77	62.07
	✓	87.60	90.16	
$K = 10$		87.37	90.08	118.92
	✓	87.71	90.34	

Table 5. The effect of the number of updates K on performance and computation costs. $K = 0$ indicates that the base model does not perform test-time adaptation. Meta-optimization enables MeTTA to achieve fast adaptation with low computation costs.

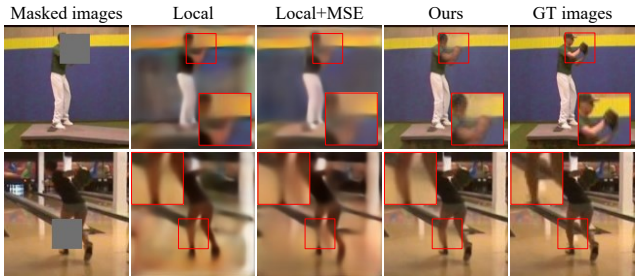


Figure 3. The effect of image inpainting under different losses. The combination of local MSE loss and global Perceptual loss [15] outputs more realistic results for the missing patches.

Number of weight updates. Limited by the multi-task learning framework, existing test-time adaptation methods [20, 38] cannot know the required number of updates that are not the same for different test images. The consequence is that insufficient updates cannot fully adapt to the target domain, and excessive updates lead to overfitting for the auxiliary task. In MeTTA, the number of updates becomes a hyper-parameter to control the execution of inter loops in meta-optimization. We demonstrate the performance and computation cost of our method with different number of updates $K = \{0, 1, 5, 10\}$ in Table 5. In order to balance accuracy and efficiency, we set the number of updates K to 5 in the rest of our experiments.

Design of the auxiliary loss. We use self-supervised image inpainting as the auxiliary task in MeTTA. To achieve fast adaptation, the implementation and optimization of the auxiliary network should not be too complicated. Under such requirements, the design of the auxiliary loss is quite important to correctly reconstruct the missing patch to learn human body information. We illustrate the output of the auxiliary network trained by different losses in Fig. 3. The experimental results show that the reconstruction quality is poor when only the local critic is used for the missing patch. The global critic for the entire image effectively reduces the difficulty of image reconstruction. Compared with the MSE loss that focuses on pixel-level differences, the Perceptual

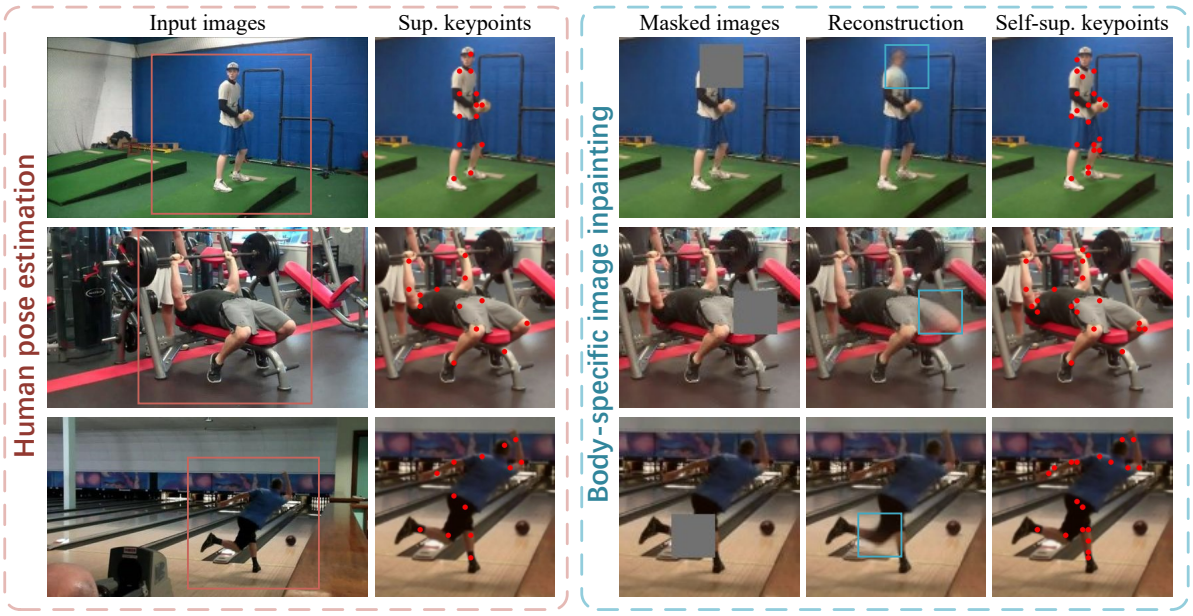


Figure 4. Visualization of MeTTA on Penn Action [48]. For the primary task, we show the input images and supervised keypoints on the left side. For the auxiliary task, we show the masked images, reconstructed images, and self-supervised keypoints on the right side.

Dataset	Mask size		
	small	middle	large
Penn Action [48]	86.80	87.60	86.05
Human3.6M [11]	89.46	90.16	89.68
MPII [1]	90.49	91.41	90.72

Table 6. Performance under different mask sizes. For the input image of 128 × 128 from Penn Action, the mask size is set to 10 × 10, 30 × 30, and 60 × 60 to reflect small, middle, and large. For the 256 × 256 input image from Human3.6M and MPII, the mask size is set proportionally to 20 × 20, 60 × 60, and 120 × 120.

loss [15] based on semantic information is more suitable for our purpose. Therefore, the training objective consists of local MSE loss and global Perceptual loss.

Different sizes of the body-specific mask. The mask size is critical for image inpainting tasks, so it affects the performance of test-time adaptation for human pose estimation. The experimental results in Table 6 show that the best performance is achieved when the mask size is set to 30 × 30 for the 128 × 128 input image and to 60 × 60 for the 256 × 256 input image. As shown in Fig 1, a part of the human body in the input image is removed, which prompts the auxiliary network to learn human-related semantics (the appearance from the remaining body pixels and the pose from the self-supervised heatmaps). If the mask size is set too small, the image inpainting task can be easily completed without the need for pose information. If the mask size is set too large, the pose information is not enough to recover the missing patch when most of the body pixels are lost.

4.6. Visualization

To better illustrate the effectiveness of MeTTA, we provide visualization on Penn Action, as shown in Fig. 4. Human pose estimation as the primary task aims to predict the 2D coordinates of body joints. Masking body parts in images, the auxiliary task is to reconstruct the missing patches based on self-supervised keypoints to correlate with the primary task. In Fig. 4, the consistency between self-supervised and supervised keypoints indicates that body-specific image inpainting indeed captures human body information.

5. Conclusion

In this paper, we propose a test-time fast adaptation method MeTTA based on meta-auxiliary learning for human pose estimation. We use novel body-specific image inpainting as the auxiliary task to focus on human-related semantics in a single image. For test-time adaptation, the introduction of meta-auxiliary learning brings two major benefits: Firstly, meta-optimization ensures that self-supervised learning of the auxiliary task facilitates human pose estimation during inference. Secondly, a good weight initialization learned from the source domain enables fast adaptation to the target data. Extensive experiments show that our method achieves superior performance on pose estimation benchmarks.

6. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NO. 62176125, 61772272).

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, 5, 6, 8

[2] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *ICCV*, pages 13359–13368, 2021. 2

[3] Zhixiang Chi, Yang Wang, Yuanhao Yu, and Jin Tang. Test-time fast adaptation for dynamic scene deblurring via meta-auxiliary learning. In *CVPR*, pages 9137–9146, 2021. 1, 2, 3

[4] Zhixiang Chi, Li Gu, Huan Liu, Yang Wang, Yuanhao Yu, and Jin Tang. MetaFSCIL: A meta-learning approach for few-shot class incremental learning. In *CVPR*, pages 14166–14175, 2022. 3

[5] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135, 2017. 3

[6] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015. 1

[7] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020. 7

[8] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 1, 2, 4, 6

[9] Liang-Yan Gui, Yu-Xiong Wang, Deva Ramanan, and Jose M. F. Moura. Few-shot human motion prediction via meta-learning. In *ECCV*, pages 432–450, 2018. 3

[10] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In *ICML*, pages 1989–1998, 2018. 1

[11] 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. 1, 5, 6, 7, 8

[12] Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *NeurIPS*, pages 2427–2440, 2021. 4

[13] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *NeurIPS*, 2018. 1, 2, 4, 6

[14] Junguang Jiang, Yifei Ji, Ximei Wang, Yufeng Liu, Jianmin Wang, and Mingsheng Long. Regressive domain adaptation for unsupervised keypoint detection. In *CVPR*, pages 6780–6789, 2021. 5, 7

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016. 4, 7, 8

[16] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 5, 6, 7

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[18] Jogendra Nath Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, and R. Venkatesh Babu. Generalize then adapt: Source-free domain adaptive semantic segmentation. In *ICCV*, pages 7046–7056, 2021. 2

[19] Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. Pose recognition with cascade transformers. In *CVPR*, pages 1944–1953, 2021. 6

[20] 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. 1, 2, 3, 4, 6, 7

[21] 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. 6

[22] Yanjie Li, Sen Yang, Peidong Liu, Shoukui Zhang, Yunxiao Wang, Zhicheng Wang, Wankou Yang, and Shu-Tao Xia. Simcc: A simple coordinate classification perspective for human pose estimation. In *ECCV*, pages 89–106, 2022. 6

[23] Huan Liu, Zijun Wu, Liangyan Li, Sadaf Salehkalaibar, Jun Chen, and Keyan Wang. Towards multi-domain single image dehazing via test-time training. In *CVPR*, pages 5831–5840, 2022. 1, 2, 3

[24] Huan Liu, Zhixiang Chi, Yuanhao Yu, Yang Wang, Jun Chen, and Jin Tang. Meta-auxiliary learning for future depth prediction in videos. In *WACV*, pages 5756–5765, 2023. 3

[25] Shikun Liu, Andrew Davison, and Edward Johns. Self-supervised generalisation with meta auxiliary learning. In *NeurIPS*, 2019. 1

[26] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *CVPR*, pages 1215–1224, 2021. 2

[27] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105, 2015. 1

[28] Weian Mao, Yongtao Ge, Chunhua Shen, Zhi Tian, Xinlong Wang, Zhibin Wang, and Anton van den Hengel. Poseur: Direct human pose regression with transformers. In *ECCV*, pages 72–88, 2022. 6

[29] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *ICCV*, pages 542–551, 2019. 2

[30] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499, 2016. 1, 2

[31] Seobin Park, Jinsu Yoo, Donghyeon Cho, Jiwon Kim, and Tae Hyun Kim. Fast adaptation to super-resolution networks via meta-learning. In *ECCV*, pages 754–769, 2020. 1

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 6

[33] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016. 4

[34] Qucheng Peng, Ce Zheng, and Chen Chen. Source-free domain adaptive human pose estimation. In *ICCV*, pages 4826–4836, 2023. 1, 2, 3, 4, 5, 7

[35] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M. Hospedales, and Tao Xiang. Incremental few-shot object detection. In *CVPR*, pages 13846–13855, 2020. 3

[36] Dripta S. Raychaudhuri, Calvin-Khang Ta, Arindam Dutta, Rohit Lal, and Amit K. Roy-Chowdhury. Prior-guided source-free domain adaptation for human pose estimation. In *ICCV*, pages 14996–15006, 2023. 2, 3, 4, 5, 6, 7

[37] 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, 4, 6

[38] 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. 1, 3, 6, 7

[39] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, pages 109–117, 2017. 5, 7

[40] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021. 2, 4

[41] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *CVPR*, pages 7201–7211, 2022. 1, 2

[42] Dong Wei, Huaijiang Sun, Bin Li, Xiaoning Sun, Shengxiang Hu, Weiqing Li, and Jianfeng Lu. NeRM: Learning neural representations for high-framerate human motion synthesis. In *ICLR*, 2024. 2

[43] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, pages 466–481, 2018. 1, 2, 4, 6

[44] Hu Yu, Jie Huang, Yajing Liu, Qi Zhu, Man Zhou, and Feng Zhao. Source-free domain adaptation for real-world image dehazing. In *ACM MM*, pages 6645–6654, 2022. 2

[45] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. HRFormer: High-resolution transformer for dense prediction. In *NeurIPS*, pages 7281–7293, 2021. 2

[46] Dan Zeng, Yuhang Huang, Qian Bao, Junjie Zhang, Chi Su, and Wu Liu. Neural architecture search for joint human parsing and pose estimation. In *ICCV*, pages 11385–11394, 2021. 2

[47] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *CVPR*, 2020. 6

[48] 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. 1, 5, 6, 8