MAPS: A Noise-Robust Progressive Learning Approach for Source-Free Domain Adaptive Keypoint Detection

Yuhe Ding, Jian Liang†, Bo Jiang, Aihua Zheng and Ran He, Senior member, IEEE

Abstract—Existing cross-domain keypoint detection methods always require accessing the source data during adaptation, which may violate the data privacy law and pose serious security concerns. Instead, this paper considers a realistic problem setting called source-free domain adaptive keypoint detection, where only the well-trained source model is provided to the target domain. For the challenging problem, we first construct a teacher-student learning baseline by stabilizing the predictions under data augmentation and network ensembles. Built on this, we further propose a unified approach, Mixup Augmentation and Progressive Selection (MAPS), to fully exploit the noisy pseudo labels of unlabeled target data during training. On the one hand, MAPS regularizes the model to favor simple linear behavior in-between the target samples via self-mixup augmentation, preventing the model from over-fitting to noisy predictions. On the other hand, MAPS employs the self-paced learning paradigm and progressively selects pseudo-labeled samples from ‘easy’ to ‘hard’ into the training process to reduce noise accumulation. Results on four keypoint detection datasets show that MAPS outperforms the baseline and achieves comparable or even better results in comparison to previous non-source-free counterparts. The code is available at https://github.com/YuheD/MAPS.

Index Terms—Source-free Domain Adaptation, Keypoint Detection, Noise-Robust Learning.

I. INTRODUCTION

Unsupervised domain adaptation (UDA), a paradigm that aims to transfer knowledge from a label-rich source domain to another unlabeled target domain, is proposed to reduce the burden of manual annotations [1]. Typically, existing UDA methods resort to feature alignment [2], [3] or domain translation [4] so that the discriminative model learned on the labeled source domain could perform well in the unlabeled target domain. However, these methods always require access to the source data when adapting to the target domain, which may be infeasible due to the growing restrictions on data privacy and storage. To tackle this issue, a line of recent works [5]–[8] consider a source-data-free domain adaptation (SFDA) scheme in which only the well-trained classification model from the source domain is provided for the target domain.

Besides the widely-studied image classification task [5], the SFDA scheme has also been extended to other visual tasks, e.g., semantic segmentation [8]–[11], object detection [12]–[14] and point cloud recognition [7], [15]. Though much effort has been devoted to the classification problem, only a few works [16] explore the SFDA scheme for regression problems.

This paper for the first time studies the SFDA scheme in a popular regression task—image keypoint detection—which aims to localize the predefined object keypoints in 2D images [17]. Prior domain adaptive keypoint detection works typically align the source and target domains by elaborately designed alignment strategies, e.g., domain adversarial training [18], [19] and style transfer model [20]. However, all of them require the existence of source data and target data at the same time. Besides, current SFDA methods for classification always enlarge the margins of boundaries between different classes on the target domain, e.g., pseudo-labeling [5], [7], information maximization [5], contrastive learning [6]. Such techniques are also not applicable to the keypoint detection task since the regression space is continuous without a clear decision boundary.

For the challenging SFDA keypoint detection task, we come up with a simple baseline method by exploiting the mean-teacher framework [21], where the teacher model is an exponential moving average (EMA) of the student network. To ensure consistency across models, target samples with different augmentations are utilized in the teacher and student models, and the mean squared error between their predictions is minimized. This method only fine-tunes the model with the consistency loss and no longer requires access to the source data. Nevertheless, it does not fully explore the semantic information in the unlabeled target domain, leading to sub-optimal adaptation performance. Therefore, we resort to pseudo-labeling [22], a popular semi-supervised learning technique that generates the pseudo labels for unlabeled data. Since there is no labeled data, solely relying on these noisy pseudo labels is inevitably harmful due to noise accumulation, misleading the model over time.

In this paper, we propose a new approach termed Mixup Augmentation and Progressive Selection (MAPS) by alleviating the noise accumulation issue for source-free keypoint detection. Firstly, MAPS mixes the target samples with their shuffle version to construct the self-mixup augmentation [23]. This manner regularizes the model to produce outputs that vary linearly with the inputs, preventing over-fitting to outliers and thereby improving the model’s robustness. Secondly, to
explore reliable pseudo labels, MAPS adopts the self-paced learning scheme (SPL) [24] to alternately select easy samples with high confidence and apply a standard regression loss on them. In fact, the number of selected samples is governed by a dynamic weight that is annealed until the training loss converges. Prioritizing learning about easy samples is proven to prevent the model from getting stuck in a bad local optimum [25]. These two strategies respectively consider the convex behavior among samples and the quality of the pseudo label per sample, making them complement each other. To validate the effectiveness of our method, we conduct three experiments on human, hand, and animal keypoint detection. The main contributions of our work are summarized as follows:

- We formulate a realistic and challenging task, source-free domain adaptive keypoint detection, which is among the first to study cross-domain regression problems without the source data.
- Built on the devised baseline via mean-teacher, we propose a new approach termed Mixup Augmentation and Progressive Selection (MAPS) which fully exploits the noisy pseudo labels in the target domain.
- Experiments on various keypoint detection benchmarks demonstrate that MAPS outperforms the baseline and achieves results comparable, or even better than, existing non-source-free counterparts.

II. RELATED WORK

A. Source-free Domain Adaptation

As a typical problem in transfer learning [1], [26], [27], unsupervised domain adaptation (UDA) [1]–[3], [28]–[31] aims to improve the performance of the model on an unlabeled target domain with the aid of label-rich source domain. UDA setting has been extensively researched in various tasks such as classification [28]–[31], segmentation [32], etc. However, as data privacy has received increasing attention in recent years, source-free domain adaptation (SFDA) [5] is proposed to solve the UDA problem with only a well-trained source model.

There are two primary types of SFDA approaches, i.e., generation-based, and self-training-based. Generation-based methods [35]–[37] attempt to restore the source features or images to compensate for the absence of source data. Self-training-based methods [5], [6], [15], [38], [39] leverage self-supervised techniques to explore the intrinsic structure of the target domain. In recent years, there are also some practices [36], [40] to solve this problem by selecting part of the target data as a pseudo source domain, to compensate for the unseen source domain. The following works develop kinds of variants of SFDA, e.g., active SFDA [41], black-box SFDA [42], multi-source SFDA [43], semi-supervised SFDA [44]. In reality, the vanilla SFDA along with these variants is mainly applied in multiple visual classification tasks, e.g., object recognition [5], semantic segmentation [8], [10], [45], and object detection [12]. The SFDA scheme for classification always enlarges the margins of decision boundaries between different classes on the target domain, thereby improving the generalization performance. However, the regression space is continuous without a clear boundary, leading to these methods cannot apply to regression tasks directly. There are a few works proposed for measurement shift [46], blind image quality assessment [16]. For the general keypoint detection task, it is still blank. To the best of our knowledge, we are the first to investigate the SFDA scheme on the keypoint detection task.

B. Domain Adaptive Keypoint Detection

Keypoint detection is a fundamental visual problem. It is laborious and time-consuming to collect the labeled data for traditional keypoint detection methods [47]–[51]. Domain adaptive keypoint detection [19], [52]–[57] attempts to address this issue by transferring information from a labeled source domain to an unlabeled target domain, avoiding the extra annotation costs. These methods are primarily divided into 3D and 2D keypoint detection. In 3D keypoint detection, this problem is complex and there are many prior works. For instance, [56] proves that the neural networks can perform well when the data is pre-processed to extract cues about the person’s motion, notably as optical flow and the motion of 2D keypoints, and therefore propose to use motion as a simple way to bridge a Synthetic-to-realistic gap when the video is available. UDA-COPE [57] introduces a bidirectional filtering method between the predicted normalized object coordinate space (NOCS) map and observed point cloud, to promote teacher-student consistency training.

This paper focuses on another type, i.e., 2D keypoint detection. In this field, existing methods can be primarily divided into human keypoint detection and animal keypoint detection. For human keypoint detection, RegDA [19], and MarsDA [58] introduce adversarial regressors to narrow the domain gap. C-GAC [52] integrates the proposed prediction confidence into self-training to obtain reliable pseudo labels. For animal keypoint detection, CC-SSL [39], and UDA-Animal [18] are based on transformation consistency and the pseudo-label refinery technique. Besides, unified domain adaptive pose estimation (UDAPE) [20] proposes to align representations using both input-level and output-level cues, and provides a unified framework for both human keypoint detection and animal keypoint detection problems. Generally, the source data is essential for these carefully designed alignment strategies, which may violate data privacy law and pose serious concerns, while our method solves this problem under a source data-free setting.

C. Curriculum Learning

Curriculum learning (CL) [60] is a training strategy that trains a model from easy to hard, which simulates the learning principle of humans and animals. CL serves as a general training strategy and has been used in multiple applications of CL in machine learning [61]–[64]. In these applications, the motivations can be categorized for applying CL into two groups: to guide, regularizing the training towards better regions in parameter space (with steeper gradients) from the perspective of the optimization problem, and to denoise, focusing on the high-confidence easier area to alleviate the interference of noisy data as from the perspective of data distribution. A general CL framework has a Difficulty Measurer,
which decides the relative “easiness” of each data example, and a Training Scheduler, which decides the sequence of data subsets throughout the training process based on the judgment from the Difficulty Measurer. Based on this framework, CL is primarily divided into two cases, i.e., predefined CL and automatic CL. The predefined CL case measures the difficulty of samples and schedules the training process based on prior knowledge, and the automatic CL case adopts a data-driven method to learn any (or both) aspects. In predefined CL, researchers manually design various Difficulty Measurers mainly based on the data characteristics of specific tasks. For instance, in NLP tasks, sentence length is the most popular Difficulty Measure in NLP tasks [65]. This is because researchers intuitively think the complexity of a sentence or paragraph can be expressed by sentence length.

Self-paced learning (SPL) [24], [66] is a typical example of automatic curriculum learning. The easiest samples with the lowest training loss are used to train the models, this portion of the easiest samples gradually increases according to a specified scheduler. Based on the scheduler, the models are able to learn at their own pace, including deciding what, how, when, and how long to study. Prior work formulates the critical principle of SPL as a concise model [24], and proves that the SPL regime is convergent by adopting an alternative optimization strategy (AOS) [67]. Generally, previous self-paced learning methods focus on the supervised learning scenario, and no few works study the application of SPL on unsupervised learning tasks. Traditional supervised SPL achieves promising results, while unsupervised SPL still remains much room for improvement. In this paper, we study unsupervised SPL in regression problems. Note that self-paced learning is a progressive learning approach, but not all progressive learning methods fall under self-paced learning. For instance, the emphasis in the progressive learning approach described in [32] lies in gradually selecting challenging samples to explore domain-specific information in the target domain. In contrast, our self-paced learning regime progressively selects easy samples to mitigate the influence of noisy labels.

### III. Method

We define the notations in Section III-A and generate the source model by a standard regression method, in Section III-B. Then we construct a simple baseline by the mean-teacher [20], [21]. Specifically, the student model \( f' \) and teacher model \( f \) have an identical architecture with the given source model \( f_s \), and the teacher parameter \( \theta' \) is updated with the exponential moving average (EMA) of the student parameter \( \theta \):

\[
\theta'_t = \eta \theta'_{t-1} + (1 - \eta) \theta_t,
\]

where \( t \) denotes the step of training, and \( \eta \) denotes the smoothing coefficient which is set to 0.999 by default.

Here we also introduce the random data augmentation function \( \mathcal{A}^t \) on target data, the input of student \( \mathcal{A}^t(x_i^t) \) and the input of teacher \( \mathcal{A}^t(x_i^t) \) are generated from different augmentation \( \mathcal{A}^t \) and \( \mathcal{A}^t \), sampled from \( \mathcal{A}^t \). Besides, considering the missing and occluded keypoints in some images, we only force the teacher-student consistency on the predicted points with the maximum activation greater than a threshold \( \tau \). The object function is formulated as:

\[
\mathcal{L}_{MT}(x_t) = \sum_{i=1}^{n_t} \sum_{k=1}^{K} \mathbb{1}(h_k^t \geq \tau) \| \mathcal{A}^t(h_k^t) - \mathcal{A}^t(h_k^t) \|_2,
\]

where \( \mathcal{A}^t \) and \( \mathcal{A}^t \) denote the inverse function of \( \mathcal{A}^t \) and \( \mathcal{A}^t \), \( \{h_k^t\}_{k=1}^{K} = f_s(\mathcal{A}^t(x_i^t)) \) denote the outputs of student model, \( \{h_k^t\}_{k=1}^{K} = H(f'_s(\mathcal{A}^t(x_i^t))) \) denote the normalized outputs of teacher model, \( \mathbb{1}(\cdot) \) denotes the indicator function, and \( K \) denotes the number of keypoints.

### A. Notations

For a standard unsupervised domain adaptive 2D keypoint detection, we have \( n_s \) labeled samples \( \{x_i^s, y_i^s\}_{i=1}^{n_s} \) from the source domain \( D_s = (X_s, Y_s^K) \), and \( n_t \) unlabeled data \( \{x_i^t\}_{i=1}^{n_t} \) from the target domain \( D_t = (X_t) \), where \( X_s, X_t \in \mathbb{R}^{H \times W \times 3} \) are the input space, \( Y_s^K \in \mathbb{R}^2 \) is the output space, \( K \) is the number of keypoints in each input image. The goal is to find a keypoint detector to predict the labels \( \{y_i^t\}_{i=1}^{n_t} \) of the target domain.

The source-free unsupervised domain adaptive keypoint detection task aims to learn a target keypoint detector \( f_t : X_t \rightarrow Y_t \) and infer \( \{y_i^t\}_{i=1}^{n_t} \) with only \( \{x_i^t\}_{i=1}^{n_t} \) and the source keypoint detector \( f_s \). We first generate the source keypoint detection model by a standard regression method, and then we transfer the model to the target domain without accessing the source data.

### B. Source Model Generation

We consider to train the source keypoint detection model \( f_s : X_s \rightarrow Y_s \) by minimizing the following standard supervised regression loss,

\[
\mathcal{L}_{src}(X_s, Y_s) = \sum_{i=1}^{n_s} \| f_s(x_i^s) - y_i^s \|_2^2,
\]

where \( H(\cdot) \) denotes the heatmap generation function [68]. To further improve the performance of the source model and facilitate the following target data alignment, we use the random data augmentation technique [18] at this stage. Taking the augmentation into consideration, the objective function is reformulated as:

\[
\mathcal{L}_{src}^{aug}(X_s, Y_s) = \sum_{i=1}^{n_s} \| f_s(A^t(x_i^s)) - A^t(H(y_i^s)) \|_2^2,
\]

where \( A^t \) denotes the augmentation function.

### C. Teacher-Student Consistency Learning

Given the well-trained source model \( f_s \), we first construct a baseline method based on the mean-teacher framework [18], [20], [21]. Specifically, the student model \( f_t \) and teacher model \( f' \) have an identical architecture with the given source model \( f_s \), and the teacher parameter \( \theta' \) is updated with the exponential moving average (EMA) of the student parameter \( \theta \):

\[
\theta'_t = \eta \theta'_{t-1} + (1 - \eta) \theta_t,
\]

where \( t \) denotes the step of training, and \( \eta \) denotes the smoothing coefficient which is set to 0.999 by default.
Fig. 1: The pipeline of Mixup Augmentation and Progressively Selection (MAPS). Given the well-trained source model $\theta_0$, MAPS progressively selects easy samples from target samples, i.e., assigns weight, 1 for easy samples and 0 for the remaining samples, the weights are notated as symbol $v$. Then these samples participate in the model updating with self-mixup training, which regularizes the model to favor a linear behavior in-between the target samples. These two aspects are alternatively performed until the loss function converges.

D. Noise-Robust Progressive Learning

Based on the baseline depicted in section III-C, we further employ a simple strategy, pseudo-labeling [22], which directly uses the current prediction of the model as the ground truth, to explore the semantic information in the target domain. However, this manner inevitably introduces many erroneous labels, making the model suffer from noise accumulation, which obstructs true signal and biasing estimation of corresponding parameters [69], leading to sub-optimal results. As shown in Fig. 1, we further propose Mixup Augmentation and Progressive Selection (MAPS), to avoid the over-confidence in outliers and select reliable pseudo labels progressively to participate in the model training.

Mixup Augmentation. Based on the baseline model, we propose self-mixup augmentation to favor simple linear behavior in-between target data during training, thereby preventing confirmation bias [23] and improving the model’s robustness. To be specific, we mix the augmented target samples $A(x_i)$ and another random sample $A'(x_i)$ in the current mini-batch, with a mix ratio $\rho \sim \text{Beta}(\alpha, \alpha)$ to construct the self-mixup augmentation, and then we introduce the self-mixup loss in the following:

$$L_{\text{mix}}(X_t) = \sum_{i=1}^{n_t} \| f_i(x_i^{m}) - h_i^{m} \|_2,$$

$$x_i^{m} = \rho A'(x_i) + (1-\rho) A(x_i),$$

$$h_i^{m} = \rho f_i(A'(x_i)) + (1-\rho) f_i(A(x_i)).$$

The self-mixup loss regularizes the model to favor simple linear behavior in-between the target samples, preventing the over-fitting of outliers.

Progressive Selection. The pseudo-labeling technique [22] picks up the current prediction as used as if they were true labels. However, it is inevitable to introduce noisy labels, which might mislead the model over time. We introduce the self-paced learning regime [24] to progressively select samples with high-confident pseudo labels, and the pseudo-label regression loss is imposed on them. This aim can be realized by minimizing the following objective:

$$L_{\text{spl}}(X_t) = \sum_{i=1}^{n_t} v_i l_{\text{reg}}(x_i^t) + R(v_i, \lambda), \text{ s.t. } v_i \in [0, 1]$$

where $\lambda$ denotes the age parameter for adjusting the learning pace, $v_i$ denotes the weight for the $i$-th sample, $R(v_i, \lambda)$ denotes the self-paced regularizer (SP-regularizer), and $l_{\text{spl}}$ denotes the self-paced loss weighted by $v_i$. The self-paced loss for each sample $x_i^t$ is defined by the mean squared error, i.e., pseudo-label regression loss:

$$l_{\text{reg}}(x_i^t) = \| f_i(A'(x_i^t)) - A'(H(\hat{y}_i^t)) \|_2,$$

where $\hat{y}_i^t$ is the pseudo label of $x_i^t$, which is the 2D coordinate of the maximum activation of the student output $f_i(x_i^t)$.

Here we adopt the hard SP-regularizer [24] to solve the sample weights. The formulation and the closed-form solution
are below:

\[ R(v_i, \lambda) = -\lambda v_i; \quad v_i = \begin{cases} 1, & l_{reg}(x_i^t) < \lambda \\ 0, & l_{reg}(x_i^t) \geq \lambda \end{cases} \quad (8) \]

The model parameters are fixed during selection, therefore, \( l_{reg} \) can be seen as a difficulty measurer (i.e., pseudo-label selection method). The age parameter \( \lambda \) is a threshold, this regime progressively selects the easy samples by assigning the weight 1 for them.

The age parameter \( \lambda \) is updated by a baby step scheduler \( S \) [70]. To be specific, we pre-define an increasing sequence \( M = \{ M_1, M_2, M_3, ..., M_Q \} \). In the \( q \)-th round, we select \( M_q \) samples, and \( \lambda \) can be calculated by:

\[ \lambda = S(M_q; l_{reg}(x_i^t))_{i=1}^{n_t}, \]

where \( S(m, q) \) denotes the \( m \)-th smallest value of loss function \( L \).

**Overall Objective.** With the above two aspects, the overall objective of MAPS can be summarized as:

\[ \min_{\theta, v} L_{mt}(\mathcal{X}_t) + \beta_m L_{mix}(\mathcal{X}_t) + \beta_s L_{spl}(\mathcal{X}_t), \]

where \( \beta_s, \beta_m \) are hyper-parameters to control the weights of the corresponding loss terms. We adopt the alternative optimization strategy to optimize the objective function in the following steps.

1) Select easy samples, i.e., solve weights \( v \), when \( \theta \) and \( \lambda \) are fixed. Optimizing Eq. (10) is equal to solving the following problem:

\[ \min_v L_{spl}(\mathcal{X}_t), \quad s.t. \quad v_i \in [0, 1], \]

we adopt the hard SP-regularizer here, and the closed-form solution of \( v \) is shown in Eq. (8).

2) Train the model, i.e., update \( \theta \), when \( v \) and \( \lambda \) are fixed. Optimizing Eq. (10) is equal to solving the following problem:

\[ \min_\theta L_{mt}(\mathcal{X}_t) + \beta_m L_{mix}(\mathcal{X}_t) + \beta_s \sum_{i=1}^{n_t} v_i l_{reg}(x_i^t). \]

In addition, according to the given baby step scheduler in Eq. (9), we also update the age parameter \( \lambda \) to include more training samples during the optimization process. The overall algorithm is shown in Algorithm 1.

**Algorithm 1** Algorithm of the proposed MAPS.

**Input:** Target dataset \( \mathcal{X}_t \); a well-trained source model \( f_s \);

**Output:** New target model \( f_t \) and \( f_t' \);

1: Initialize the student \( f_t \) and teacher \( f_t' \) with \( f_s \);
2: repeat
3: Fix \( f_t \), update \( v \) by Eq. (11); // sample selection
4: Update \( f_t \) by Eq. (12); // model updating
5: Update \( f_t' \) by Eq. (3);
6: Update \( \lambda \) based on \( S \) by Eq. (9);
7: until Progressive selection rounds are exhausted.

**IV. EXPERIMENTS**

We first give the experimental setup in Section IV-A, then we conduct the experiments and analyze the results on hand keypoint detection, human keypoint detection, and animal keypoint detection in Section IV-B, IV-C, IV-D, respectively. To validate the effectiveness of the proposed modules, we conduct the ablation study in Section IV-E and analyze the sensitivity in Section IV-F.

**A. Setup**

The architecture of our model is based on Simple Baseline [49] with the backbone of pre-trained ResNet101 [71]. Following prior works [18], [20], [59], we add texture and geometric augmentation including Gaussian noise, Gaussian blurring, rotation, and random 2D translation for both source model generation and target model training. The parameter \( \alpha \) of the beta distribution is set to 0.75, and the weights of loss functions \( \beta_m \) and \( \beta_s \) are both empirically set to 1. In the source model generation stage, we adopt Adam [72] as the optimizer and set the learning rate as 2e-4, and it decreased to 2e-5 at 15,000 steps and 2e-6 at 20,000 steps, while there is a total of 25,000 steps in this stage. In the target model training, we also adopt the Adam optimizer and learning rate 2e-4, and the learning rate decreases to 1e-4 at 2,500 steps with a total of 7,500 steps. Our basic augmentation is based on UDAPE [20], which adds rotation and random 2D translation on the augmentation in RegDA [19]. We use the Percentage of Correct Keypoints (PCK) as our metric, in which estimation is considered correct if its distance from the ground truth is less than a fraction of 0.05 of the image size. We randomly run our methods three times with different random seeds \( \{0,1,2\} \) via PyTorch and report the average accuracies.

We conduct experiments on three tasks with four different target datasets, and compare MAPS with several state-of-the-art non-source-free methods, including the animal keypoint detection methods CCSSL [59] and UDA-Animal [18]; the human and hand keypoint detection method RegDA [19], and the unified method UDAPE [20]. In the table of comparative experiments, ‘source only’ denotes using the source model for prediction, ‘MT’ is our baseline method introduced in section III-C, and MAPS is the full version of our method. ‘Oracle’ corresponds to training on the target domain with supervised data. Specifically, given supervised target domain data, we utilize a standard ResNet101 network with a regression head and employ the standard mean squared error (MSE) loss for training in order to obtain the Oracle results. This represents the theoretically optimal performance that can be achieved for this task. All the results except ours are from the original paper of UDAPE, and the dataset split policy follows UDAPE and RegDA.

**B. Hand Keypoint Detection**

**Dataset.** Rendered Hand Pose Dataset [73] (RHD) is a synthetic dataset containing 41,258 training images and 2,728 testing images with corresponding 21 hand keypoints labels. Hand-3D-Studio [74] (H3D) is a real-world multi-view hand...
image dataset containing 22k images in total. Following the split way in RegDA [19], we select 3.2k images as the testing set, and the remaining as the training set.

### Implementation Details
We use RHD as the source domain and H3D as the target domain. The number of progressive selection rounds \( Q \) and the increasing sequence \( M \) in the baby step scheduler are set to 3 and \( \{0.25N, 0.35N, 0.45N\} \), respectively, where \( N \) is the total number of training samples. We report 21 keypoints on the different anatomical parts of a hand including metacarpophalangeal (MCP), proximal interphalangeal (PIP), distal interphalangeal (DIP), and fingertip (Fin).

### Results
The quantitative results are presented in Table I. In comparison, MAPS achieves the highest average accuracy even compared with the non-source-free methods, and MT also achieves competitive results. We outperform a large margin than most of the methods except UDAPE [20], demonstrating the effectiveness of our method. MAPS also outperforms MT, as the noise accumulation phenomenon in MT is reduced. We show the qualitative results in Fig. 2. The results of ‘source-only’ are not satisfied in some complicated hand poses due to a large domain gap, and MT successfully aligns the target domain to the unseen source domain. MAPS further refines the results of MT, obtaining more accurate results.

### Human Keypoint Detection
**Dataset.** SURREAL [75] is a synthetic dataset that consists of monocular videos of people with a total of 6M images, which are photo-realistic renderings of people under large variations in shape, texture, viewpoint, and pose. Leeds Sports Pose [76] (LSP) is widely used as the benchmark for human keypoint detection. LSP contains a total of 2k images of sports persons with annotated human body joint locations gathered from Flickr.

**Implementation Details.** We use SURREAL as the source domain and LSP as the target domain. The increasing sequence \( M \) in the baby step scheduler is set to \( \{0.1N, 0.2N, 0.3N\} \), where \( N \) is the total number of training samples, and the number of progressive selection rounds \( Q \) is 3. We report 16 keypoints on the human body parts, i.e., shoulder (Sld), elbow (Elb), wrist, hip, knee, and ankle.

### Animal Keypoint Detection
**Dataset.** Synthetic Animal Dataset [59] (SynAnimal) is a synthetic pose dataset rendered from CAD models, containing 5 different animals including the horse, tiger, sheep, hound,
### TABLE II: PCK@0.05 on task SURREAL $\rightarrow$ LSP.

<table>
<thead>
<tr>
<th>Method</th>
<th>SF</th>
<th>Sld</th>
<th>Elb</th>
<th>Wrist</th>
<th>Hip</th>
<th>Knee</th>
<th>Ankle</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSSL [59]</td>
<td>×</td>
<td>36.8</td>
<td>66.3</td>
<td>63.9</td>
<td>59.6</td>
<td>67.3</td>
<td>70.4</td>
<td>60.7</td>
</tr>
<tr>
<td>UDA-Animal [18]</td>
<td>×</td>
<td>61.4</td>
<td>77.7</td>
<td>75.5</td>
<td>65.8</td>
<td>76.7</td>
<td>78.3</td>
<td>69.2</td>
</tr>
<tr>
<td>RegDA [19]</td>
<td>×</td>
<td>62.7</td>
<td>76.7</td>
<td>71.1</td>
<td>81.0</td>
<td>80.3</td>
<td>75.3</td>
<td>74.6</td>
</tr>
<tr>
<td>UDAPE [20]</td>
<td>×</td>
<td>69.2</td>
<td>84.9</td>
<td>83.3</td>
<td>85.5</td>
<td>84.7</td>
<td>84.3</td>
<td>82.0</td>
</tr>
</tbody>
</table>

| Source only     | -  | 50.6| 64.8| 63.3  | 70.1 | 71.2 | 70.1  | 65.0 |
| MT              | ✓  | 69.2| 82.1| 80.2  | 83.4 | 82.3 | 80.8  | 79.6 |
| MAPS            | ✓  | 67.0| 84.2| 82.7  | 84.0 | 83.8 | 84.6  | 81.0 |

| Oracle          | -  | 100.0| 100.0| 100.0 | 100.0| 100.0| 100.0 | 100.0 |

### TABLE III: PCK@0.05 on task SynAnimal $\rightarrow$ AnimalPose.

<table>
<thead>
<tr>
<th>Method</th>
<th>SF</th>
<th>Dog</th>
<th>Sheep</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eye Hoof Knee Elb</td>
<td></td>
<td>Avg.</td>
<td></td>
</tr>
<tr>
<td>Eye Hoof Knee Elb</td>
<td></td>
<td>Avg.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>SF</th>
<th>Dog</th>
<th>Sheep</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eye Hoof Knee Elb</td>
<td></td>
<td>Avg.</td>
<td></td>
</tr>
<tr>
<td>Eye Hoof Knee Elb</td>
<td></td>
<td>Avg.</td>
<td></td>
</tr>
</tbody>
</table>

| Source only     | -  | 50.6| 64.8| 63.3  | 70.1 | 71.2 | 70.1  | 65.0 |
| MT              | ✓  | 69.2| 82.1| 80.2  | 83.4 | 82.3 | 80.8  | 79.6 |
| MAPS            | ✓  | 67.0| 84.2| 82.7  | 84.0 | 83.8 | 84.6  | 81.0 |

| Oracle          | -  | 100.0| 100.0| 100.0 | 100.0| 100.0| 100.0 | 100.0 |

### TABLE IV: PCK@0.05 on task SynAnimal $\rightarrow$ TigDog.

<table>
<thead>
<tr>
<th>Method</th>
<th>SF</th>
<th>Horse</th>
<th>Tiger</th>
</tr>
</thead>
</table>

| Source only     | -  | 50.6| 64.8| 63.3  | 70.1 | 71.2 | 70.1  | 65.0 |
| MT              | ✓  | 69.2| 82.1| 80.2  | 83.4 | 82.3 | 80.8  | 79.6 |
| MAPS            | ✓  | 67.0| 84.2| 82.7  | 84.0 | 83.8 | 84.6  | 81.0 |

| Oracle          | -  | 100.0| 100.0| 100.0 | 100.0| 100.0| 100.0 | 100.0 |

### TABLE V: The effectiveness of three loss functions and their interactions on task SynAnimal $\rightarrow$ AnimalPose.

<table>
<thead>
<tr>
<th>Loss Function</th>
<th>SynAnimal $\rightarrow$ AnimalPose: Dog</th>
<th>SynAnimal $\rightarrow$ AnimalPose: Sheep</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{L}_{mt}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mathcal{L}_{mix}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mathcal{L}_{spl}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Eye Hoof Knee Elb | Avg. | Eye Hoof Knee Elb | Avg. |
| Eye Hoof Knee Elb | Avg. | Eye Hoof Knee Elb | Avg. |

### TABLE VI: The effectiveness of three loss functions and their interactions on task SURREAL $\rightarrow$ LSP.

<table>
<thead>
<tr>
<th>Loss Function</th>
<th>SURREAL $\rightarrow$ LSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{L}_{mt}$</td>
<td>Sld Elb Wrist Hip Knee Ankle</td>
</tr>
<tr>
<td>$\mathcal{L}_{mix}$</td>
<td></td>
</tr>
<tr>
<td>$\mathcal{L}_{spl}$</td>
<td></td>
</tr>
</tbody>
</table>

| Source only     | -  | 50.6| 64.8| 63.3  | 70.1 | 71.2 | 70.1  | 65.0 |
| MT              | ✓  | 69.2| 82.1| 80.2  | 83.4 | 82.3 | 80.8  | 79.6 |
| MAPS            | ✓  | 67.4| 83.3| 80.5  | 83.9 | 83.6 | 83.7  | 80.4 |

| Source only     | -  | 67.0| 84.2| 82.7  | 84.0 | 83.8 | 84.6  | 81.0 |

| Source only     | -  | 67.0| 84.2| 82.7  | 84.0 | 83.8 | 84.6  | 81.0 |

| Source only     | -  | 67.0| 84.2| 82.7  | 84.0 | 83.8 | 84.6  | 81.0 |
Fig. 3: Qualitative results of task SURREAL → LSP.

Fig. 4: Qualitative results of task SynAnimal → AnimalPose.

and elephant. Each class contains 10k images. AnimalPose dataset [54] contains 6.1k images in the wild from 5 animals, i.e., dog, cat, cow, sheep, and horse. TigDog dataset [77] is a large-scale animal keypoint detection dataset, which provides 30k images from real-world videos of horse and tigers.

Implementation Details. We conduct two experiments in this section, with SynAnimal as the source domain in both experiments, AnimalPose and TigDog as the target domain, respectively. In SynAnimal → AnimalPose, the increasing sequence $M$ is set to \{0.3N, 0.4N, 0.5N\}, and the number of progressive selection rounds $Q$ is set to 3. We report the detection results of 14 keypoints on the dog and sheep body parts including the eye, hoof, knee, and elbow (Elb). In SynAnimal → TigDog, $M$ is set to \{0.1N, 0.15N, 0.2N\}, and $Q$ is also set to 3. We present the detection results of 14 keypoints on the horse and tiger body parts including the eye, chin, shoulder (Sld), hip, elbow (Elb), knee, and hoof. We run these experiments with the same augmentation in UPADE.

Results The results of SynAnimal → AnimalPose are shown in Table III. MAPS achieves the best results among these methods on both dog and sheep keypoint detection and improves the baseline method MT by a large margin (2%). AnimalPose is a small dataset, which makes the model easily overfitting to outliers, and our self-mixup loss favors the predictions to vary linearly with the input to improve the robustness, therefore performing well on such a small dataset. The qualitative results are shown in Fig. 4. The results of SynAnimal → TigDog are shown in Table IV. In this task, MAPS still has room for improvement. The reason why MAPS performs not well is probably due to the fact that the domain gap between SynAnimal and TigDog is primarily reflected in image style. UDAPE [20] directly solves the style alignment by introducing
TABLE VII: Ablation study on different sample selection strategies on task SynAnimal→TigDog. ‘Full set’: selecting the full target set as easy samples in one round. ‘One round’: selecting part of confident samples in one round. ‘SPL’: selecting part of samples by self-paced learning strategy of MAPS.

<table>
<thead>
<tr>
<th>Selection</th>
<th>SynAnimal→TigDog: Dog</th>
<th>SynAnimal→TigDog: Sheep</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Eye</td>
<td>Hoof</td>
</tr>
<tr>
<td>Full set</td>
<td>58.5</td>
<td>60.7</td>
</tr>
<tr>
<td>One round</td>
<td>61.9</td>
<td>61.3</td>
</tr>
<tr>
<td>SPL</td>
<td>63.2</td>
<td>60.5</td>
</tr>
</tbody>
</table>

TABLE VIII: Ablation study on different sample selection strategies on task SURREAL→LSP.

<table>
<thead>
<tr>
<th>Selection</th>
<th>SURREAL→LSP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sld</td>
</tr>
<tr>
<td>Full set</td>
<td>54.3</td>
</tr>
<tr>
<td>One round</td>
<td>65.3</td>
</tr>
<tr>
<td>SPL</td>
<td>64.4</td>
</tr>
</tbody>
</table>

(a) Influence of the weight $\beta_m$. (b) Influence of the weight $\beta_s$. (c) Influence of the rounds $Q$.

Fig. 5: Analysis of the weights $\beta_m$ and $\beta_s$, and the number of progressive selection rounds $Q$ on task SURREAL→LSP.

Effectiveness of sample selection strategies. We conduct an ablation study on different sample selection methods. The results are presented in Table VII and VIII. In the tables, ‘Full set’ denotes selecting the full target set as easy samples in one round. ‘One round’ denotes selecting part of confident samples in one round. This can be seen as a special case of progressive selection, where $Q = 1$. ‘SPL’ denotes the progressive selection strategy in MAPS, i.e., selecting part of samples by self-paced learning strategy. By on both two tasks, the ‘SPL’ row obtains the highest accuracy, which demonstrates that selecting the part of confident samples in one round filters out most of the noisy pseudo labels, and the performance is improved. On task SynAnimal→AnimalPose, selecting the full target set as easy samples is better than selecting part of confident samples in one round. Although the full target set has more noisy labels, it also contains a larger amount of data and more information. The benefits of the large data and information can compensate for the negative effects of noise. On task SURREAL→LSP, the opposite is true. The negative effect of noise from using the full dataset is severe, and selecting part of confident samples alleviates this issue. In fact, this is a trade-off problem between noise and information, our progressive selection strategy improves the pseudo-label quality by leveraging the dynamic capacity of the model to achieve the best results.
This paper considers a realistic setting called source-free domain adaptive keypoint detection, where only the well-trained source model is provided to the target domain. We first construct a simple baseline method based on the mean-teacher framework. Then we propose a new approach termed Mixup Augmentation and Progressive Selection (MAPS) built on this. The mixup augmentation regularizes the model to favor simple linear behavior in-between target samples thereby improving the robustness. The progressive selection strategy leverages the dynamic capacity of the current model to explore reliable pseudo labels. These two strategies respectively consider the convex behavior among samples and the quality of the pseudo label per sample, making them complement each other. Experiments verify that MAPS achieves competitive and even state-of-the-art performance in comparison to previous non-source-free counterparts.

VI. ACKNOWLEDGEMENTS

The authors would like to thank the reviewers and the associate editor for their valuable comments. The authors also thank Lijun Sheng for his help in improving the technical writing aspect of this paper. This work is funded by the National Natural Science Foundation of China (62076004, 62276256), Beijing Nova Program (Z211100002121108), and Natural Science Foundation of Anhui Province (2108085Y23).

REFERENCES

Yuhe Ding is currently a Ph.D. student in the School of Computer Science and Technology, Anhui University, Hefei, China. She received her B.Eng. degree from Anhui University in July 2019. From 2020 to 2022, she was a visiting student at CASIA. Her research interests focus on transfer learning, pattern recognition, and computer vision.

Jian Liang (Member, IEEE) received the B.E. degree in Electronic Information and Technology from Xi'an Jiaotong University and Ph.D. degree in Pattern Recognition and Intelligent Systems from NLPR, CASIA in July 2013, and January 2019, respectively. He was a research fellow at the National University of Singapore from June 2019 to April 2021. Now he joins NLPR as an associate professor. His research interests focus on transfer learning, pattern recognition, and computer vision.

Bo Jiang received the B.S. degrees in mathematics and applied mathematics and the M.Eng. and Ph.D. degree in computer science from Anhui University of China in 2009, 2012, and 2015, respectively. He is currently an associate professor in computer science at Anhui University. His current research interests include image matching, graph data representation and learning and few-shot learning. He has published more than 80 papers including 30 papers in top conference and journal such as CVPR, NeurIPS, IJCV, IEEE TPAMI, IEEE TIP and IEEE TNNLS. He gained some research projects of NSFC and Natural Science Foundation of Anhui Province. He also received the rising star award of ACM Hefei.

Aihua Zheng received B.Eng. degrees and finished Master-Doctor combined program in Computer Science and Technology from Anhui University of China in 2006 and 2008, respectively. And received Ph.D. degree in computer science from University of Greenwich of UK in 2012. She visited University of Stirling and Texas State University during June to September in 2013 and during September 2019 to August 2020 respectively. She is currently an Associate Professor and PhD supervisor at the School of Artificial Intelligence, Anhui University. Her main research interests include vision based artificial intelligence and pattern recognition, especially on person/vehicle re-identification, audio visual computing, and multimodal intelligence.

Ran He received the B.E. degree in Computer Science from Dalian University of Technology, the M.S. degree in Computer Science from Dalian University of Technology, and Ph.D. degree in Pattern Recognition and Intelligent Systems from CASIA in 2001, 2004 and 2009, respectively. Since September 2010, Dr. He has joined NLPR where he is currently a full Professor. He serves as an associate editor of Neurocomputing (Elsevier), and serves on the program committee of several conferences. His research interests focus on information theoretic learning, pattern recognition, and computer vision.