Knowledge as Priors: Cross-Modal Knowledge Generalization for Datasets without Superior Knowledge

Long Zhao\(^1\) Xi Peng\(^2\) Yuxiao Chen\(^1\) Mubbasir Kapadia\(^1\) Dimitris N. Metaxas\(^1\)
\(^1\)Rutgers University \(^2\)University of Delaware
{lz311,yc984,mk1353,dnm}@cs.rutgers.edu, xipeng@udel.edu

Abstract

Cross-modal knowledge distillation deals with transferring knowledge from a model trained with superior modalities (Teacher) to another model trained with weak modalities (Student). Existing approaches require paired training examples exist in both modalities. However, accessing the data from superior modalities may not always be feasible. For example, in the case of 3D hand pose estimation, depth maps, point clouds, or stereo images usually capture better hand structures than RGB images, but most of them are expensive to be collected. In this paper, we propose a novel scheme to train the Student in a Target dataset where the Teacher is unavailable. Our key idea is to generalize the distilled cross-modal knowledge learned from a Source dataset, which contains paired examples from both modalities, to the Target dataset by modeling knowledge as priors on parameters of the Student. We name our method “Cross-Modal Knowledge Generalization” and demonstrate that our scheme results in competitive performance for 3D hand pose estimation on standard benchmark datasets.

1. Introduction

Leveraging multi-modal knowledge to boost the performance of classic computer vision problems, such as classification [28, 35, 49], object detection [14, 39, 50] and gesture recognition [1, 7, 40, 44, 53, 57, 58], has emerged as a promising research field in recent years. Current paradigms for transferring knowledge across modalities involve aligning feature representations from multiple modalities of data during training, and then improving the performance of a unimodal system during testing with the aligned feature representations. Several different schemes for learning these feature representations have been proposed over the years [1, 40, 48, 49], and all of these rely on the availability of paired training samples from different modalities.

Recently, Gupta et al. [14] have introduced Cross-Modal Knowledge Distillation (CMKD) which is a generic yet efficient scheme among these. They transfer knowledge across different modalities by a Teacher-Student scheme [16, 41, 54]. Generally, teacher networks deliver excellent performance since they are trained on modalities with superior knowledge. However, data of these modalities may be limited or expensive to be collected. On the other hand, a student network is trained using a weak modality and thereby often results in lower performance. The goal of knowledge distillation is to transfer superior knowledge from teachers to the student by aligning their intermediate feature representations. For simplicity, in this paper, we consider a form of cross-modal knowledge distillation problems in datasets where only two modalities, i.e., one teacher and one student, are involved as shown in Fig. 1 (a).

The question we ask in this work is, what is the analogue of this paradigm for datasets which do not have modalities with superior knowledge? As a motivating exam-
ple, consider the case of 3D hand pose estimation. There are a number of “superior” modalities beyond RGB images which capture more accurate 3D hand structures, e.g., depth maps [32, 45, 59], point clouds [13, 23], or stereo images [55]. These data together with their paired RGB images can be collected by corresponding devices or synthesized using pre-defined hand shape models [12, 37]. However, most of real-world datasets still come with only a single weak modality, i.e., RGB images, which raises the question: is it possible for neural networks to transfer learned cross-modal knowledge to those target datasets where superior modalities are absent?

We answer this question in this paper and propose a technique to transfer learned cross-modal knowledge from a source dataset, where both modalities are available, to the target dataset, where only one weak modality exists. Our technique uses “paired” data from the two modalities in the source dataset to distill cross-modal knowledge, and leverages meta-learning to generalize the knowledge to the target dataset by treating it as priors on the parameters of the student network. We call our scheme Cross-Modal Knowledge Generalization (CMKG), which is illustrated in Fig. 1 (b). We further evaluate the performance of the proposed scheme in 3D hand pose estimation. We show that our generalized knowledge serves as a good regularizer to help the network learn better representations for 3D hands, and improves final results in the target dataset as well.

Our work makes the following contributions. First, unlike existing methods that distill knowledge across modalities in the same dataset, we introduce a novel method for Cross-Modal Knowledge Generalization, which generalizes the learned knowledge in the source to a target dataset where the superior modality is unavailable. Second, we introduce a novel meta-learning approach to transfer knowledge across datasets. Specifically, in Sect. 3, a simple yet powerful method is presented to distill cross-modal knowledge in the source dataset. The learned knowledge in the source dataset is then regarded as priors on network parameters during the training procedure in the target dataset. Sect. 4 describes the meta-learning algorithm for learning these priors. Third, we comprehensively evaluate our scheme in 3D hand pose estimation and demonstrate its comparable performance to the state-of-the-art methods in Sect. 5. Note that our scheme can be easily generalized to different tasks, and we leave this for future work.

2. Related Work

Knowledge Distillation. The concept of knowledge distillation was first shown by Hinton et al. [16]. Subsequent research [2, 6, 36] enhanced distillation by matching intermediate representations in the networks along with outputs using different approaches. Zagoruyko and Komodakis [54] proposed to align attentional activation maps between networks. Srinivas and Fleuret [41] improved it by applying Jacobian matching to networks. Recently, cross-modal knowledge distillation [14, 49, 53] extended knowledge distillation by applying it to transferring knowledge across different modalities. Our approach generalizes cross-modal knowledge distillation to target datasets where superior modalities are missing.

Meta-Learning. Meta-learning is also known as “learning to learn”, which intends to learn how learning can be performed in a more efficient manner. Previous approaches studied this problem from a probabilistic modeling perspective [8, 21] or in metric spaces [25, 29, 38]. Recent remarkable advances in gradient-based optimization approaches have rekindled the interest in meta-learning. Among these, Model-Agnostic Meta-Learning (MAML) [9] is proposed to solve few-shot learning. Li et al. [22] extended MAML for domain generalization. Balaji et al. [3] introduced a meta-regularization function to train networks which can be easily generalized to different domains. Our meta-learning algorithm follows the spirit of these gradient-based methods but aims to learn cross-modal knowledge as priors.

3D Hand Pose Estimation. Estimating 3D hand poses from depth maps has made great progress in the past few years [10, 11, 26, 32, 43]. On the other hand, 3D hand pose estimation from RGB images is significantly more challenging. Zimmermann and Brox [59] first proposed a deep network to learn a network-implicit 3D articulation prior together with 2D key points for predicting 3D hand poses. Other studies [40, 51, 52] learned latent representations with a variational auto-encoder for inference of 3D hand poses. Note that some recent methods [4, 12, 31, 56] focused on recovering the full shapes of 3D hands other than locations of key hand joints, which have a different research target compared with our work.

Yuan et al. [53] is the most related work in spirit to ours. Like our work, they employed cross-modal knowledge distillation to improve the performance of RGB-based 3D hand pose estimation. Our method differs significantly in that in addition to knowledge distillation, we aim to address a more challenging problem of transferring cross-modal knowledge to target datasets where depth maps are unavailable.

3. Cross-Modal Knowledge Distillation

We assume that the input data is available in two modalities $x_i$ and $\tilde{x}_i$, where $x_i$ owns superior knowledge than $\tilde{x}_i$. For each modality, one network is primarily trained with the data from its own modality. To be specific, we train a teacher network $g$ using $x_i$ and a student network $f$ using $\tilde{x}_i$. Given the ground truth $y_i$, the teacher network parameterized by $\psi$ minimizes the following $\ell^2$ regression loss:

$$L_{\text{REG}}(\tilde{x}_i; y_i; \psi) = \| g(\tilde{x}_i; \psi) - y_i \|^2. \quad (1)$$

During the training of the student network, the goal of
cross-modal knowledge distillation is to improve the learning process by transferring the knowledge from the teacher to student. The transferred knowledge can be viewed as an extra supervision in addition to the ground truth. To this end, the knowledge of networks is shared by aligning the semantics of the deep representations, i.e., activation maps of intermediate layers, between the teacher and student.

Let \( Q_j \in \mathbb{R}^{C \times H \times W} \) denote the activation map of the \( j \)-th layer in the network, which consists of \( C \) feature channels with spatial dimensions \( H \times W \). We feed \( x_i \) to the student network \( f \) and its paired \( \tilde{x}_i \) to the pre-trained teacher network \( g \). Their last activation maps \( Q_i \) are aligned by:

\[
\mathcal{L}_{\text{ACT}}(x_i, \tilde{x}_i; \theta) = \| Q_i(x_i; f) - Q_i(\tilde{x}_i; g) \|^2, \tag{2}
\]

where \( \theta \) are the parameters of the student network. Furthermore, we also match the attention maps \([54]\) of the intermediate layers between the teacher and student. Specifically, let \( A_j \in \mathbb{R}^{H \times W} \) be the channel-wise attention map of \( Q_j \) calculated by \( A_j = \sum_{i=1}^{C} \| Q_{j}^{(i)} \|^2 \), where \( Q_{j}^{(i)} \) represents the \( i \)-th channel of \( Q_j \). Then \( A_j \) is \( \ell^2 \)-normalized using \( \tilde{A}_j = A_j / \| A_j \| \), and we define the attention loss as:

\[
\mathcal{L}_{\text{ATT}}(x_i, \tilde{x}_i; \theta) = \sum_{i \in I} \| \tilde{A}_i(x_i; f) - \tilde{A}_i(\tilde{x}_i; g) \|^2, \tag{3}
\]

where \( I \) denote the indices of all teacher-student activation layer pairs for which we want to transfer attention maps. Our full knowledge distillation loss can be written as:

\[
\mathcal{L}_{\text{DIST}}(x_i, \tilde{x}_i; \theta) = \mathcal{L}_{\text{ACT}} + \lambda \cdot \mathcal{L}_{\text{ATT}}, \tag{4}
\]

where \( \lambda \) is a hyper-parameter which is set to \( 1.0 \times 10^3 \) empirically in the rest of the paper. The final student network is trained with the regression loss \( \mathcal{L}_{\text{REG}} \) in Eq. (1) together with the distillation loss \( \mathcal{L}_{\text{DIST}} \) in Eq. (4). The whole pipeline of our approach is summarized in Fig. 2.

### 4. Cross-Modal Knowledge Generalization

Consider two datasets: \( D_S = \{x_i^S, y_i^S\} \) is a source dataset while \( D_T = \{x_i^T, y_i^T\} \) denotes a target dataset. Cross-modal knowledge can be efficiently distilled in the source dataset by neural networks as shown in Sect. 3, since training pairs \((x_i^S, \tilde{x}_i^S)\) are available in \( D_S \). However, due to the absence of \( \mathcal{K}_T = \{\tilde{x}_i^T\} \), direct knowledge distillation is impossible in the target dataset \( D_T \).

In this paper, we address a novel and challenging task of **Cross-Modal Knowledge Generalization**. Specifically, we aim to learn the network parameters \( \theta_{\text{DIST}} \) which contain superior knowledge \( \mathcal{K}_T \) in the target dataset \( D_T \). As mentioned above, the main challenge is that \( \mathcal{K}_T \) is unavailable in \( D_T \). Our key idea is to generalize the learned knowledge from \( D_S \) to \( D_T \). This is achieved by interpreting knowledge as priors on the network parameters, which can be learned in \( D_S \) with meta-learning. In the following sections, we first derive our formulation from a probabilistic view. Then we present the meta-learning algorithm for knowledge generalization and theoretically show its connection to the expectation maximization (EM) algorithm.

#### 4.1. Knowledge as Priors

From a Bayesian perspective, a neural network can be viewed as a probabilistic model \( P(y_i|x_i; \theta) \): given an input \( x_i \), the network assigns a probability to each possible \( y_i \in \mathcal{Y} \) with the parameters \( \theta \). Here, we consider a regression problem where \( P(y_i|x_i; \theta) \) is a Gaussian distribution which corresponds to a mean squared loss, and \( x_i \) is mapped onto the parameters of a distribution on \( \mathcal{Y} \) using network layers parameterized by \( \theta \). Given a dataset \( \mathcal{D} = \{x_i, y_i\}_i \), \( \theta \) can be learned by maximum likelihood estimation (MLE):

\[
\theta_{\text{MLE}} = \arg \max_{\theta} \log P(\mathcal{D} | \theta) = \arg \max_{\theta} \sum_i \log P(y_i | x_i; \theta). \tag{5}
\]

We assume that \( \log P(\mathcal{D} | \theta) \) is differentiable w.r.t. \( \theta \), and then Eq. (5) is typically solved by gradient descent.

The objective of cross-modal knowledge generalization is to find the parameters \( \theta_{\text{DIST}} \) by using the training examples in \( D_T \) with intractable knowledge \( \mathcal{K}_T \). This leads to maximize the posterior density of the parameters \( \theta \) directly depends on \( D_T \) and implicitly depends on \( \mathcal{K}_T \). In order to explicitly capture this dependence, we introduce a latent...
variable $\phi$ summarizing the knowledge carried by $K_T$:

$$P(\theta|D_T, K_T) = \int P(\theta, \phi|D_T, K_T) d\phi$$

$$= \int P(\theta|D_T, K_T, \phi)P(\phi|D_T, K_T) d\phi$$

$$= \int P(\theta|D_T, \phi)P(\phi|D_T, K_T) d\phi.$$  

(6)

Note that the last equation is the result of assuming that $K_T$ and $\theta$ are conditionally independent given the latent variable $\phi$. Since both $K_T$ and integrating Eq. (6) over $\phi$ are intractable, we make an approximation that uses a point estimation $\phi_{\text{META}}$. This point estimation is obtained via the meta-learning approach described in Sect. 4.2, hence avoiding the need to perform integration over $\phi$ or interact $K_T$. Consequently, maximizing the logarithm of the posterior density of Eq. (6) can be written as:

$$\theta_{\text{DIST}} = \arg\max_{\theta} \log P(\theta|D_T, K_T)$$

$$\approx \arg\max_{\theta} \log P(\theta|D_T, \phi_{\text{META}})$$

$$= \arg\max_{\theta} \left[ \underbrace{\log P(D_T|\theta)}_{\text{Likelihood}} + \underbrace{\log P(\theta|\phi_{\text{META}})}_{\text{Prior (Knowledge)}} \right].$$  

(7)

where the last equality results from a direct application of Bayes rule. So, finding the parameters $\theta_{\text{DIST}}$ involves a two step training procedure: (1) optimizing the prior term which obtains the point estimation $\phi_{\text{META}}$ using meta-learning and (2) optimizing the likelihood term which maximizes Eq. (7) using the learned parameters $\phi_{\text{META}}$.

In a Bayesian setting, priors on the parameters can be interpreted as regularization. Thus the prior term in Eq. (7) is implemented as a regularizer during network training. Several other regularization schemes have been proposed in the literature such as weight decay [20], dropout [42, 47] and batch normalization [17]. While they aim to reduce error on examples drawn from the test distribution, the objective of our work is to learn a regularizer that captures cross-modal knowledge learned from the source dataset.

4.2. Learning Priors with Meta-Learning

As mentioned above, we model the prior term as a regularizer $R(\theta, \phi)$. Given the input $\theta$, $R$ is implemented with a neural network parameterized by $\phi$.

As described in Sect. 3, cross-modal knowledge distillation leads to optimizing the following objective:

$$G(x_i, \hat{x}_i, y_i; \theta) = L_{\text{REG}}(x_i, y_i; \theta) + L_{\text{DIST}}(x_i, \hat{x}_i; \theta),$$  

(8)

where $L_{\text{REG}}$ is the regression loss minimizing the mean squared errors of the prediction and ground truth, and the distillation loss $L_{\text{DIST}}$ distills knowledge from the teacher to student. Using the regularizer $R$, we introduce a regularized regression loss which is defined as:

$$F(x_i, y_i; \theta, \phi) = L_{\text{REG}}(x_i, y_i; \theta) + R(\theta, \phi).$$  

(9)

During the training procedure on the source dataset, we aim to learn the regularizer $R$ in Eq. (9) which mimics the behavior of $L_{\text{DIST}}$ in Eq. (8), so that $F$ can be applied to the target dataset where the superior knowledge is missing. We now describe the procedure for learning $R$.

When training the student network on the source dataset, at iteration $k$, we begin by sampling a mini-batch from the dataset. Using this batch, $l$ steps of gradient descent are first performed with the regularized regression loss $F$. Let $\theta_k$ denote the network parameters after these $l$ steps. Then the full loss $G$ on the same batch computed using $\theta_k$ is minimized w.r.t. the regularizer parameters $\phi$. The regularizer $R$ is finally updated with the gradients which unroll through the $l$ gradient steps. This ensures that $G$ can be approximated by $F$ using $R$. After finishing the training, since the same regularizer is trained on every pair of $(x_i, \hat{x}_i)$, the resulting $R$ captures the notion of cross-modal knowledge contained in the source dataset. Please refer to Fig. 3 for an illustration of the meta-training step. The entire algorithm is given in Algorithm 1. Note that $l$ is set to 1 empirically in this paper, as we observe that a 1-step update is sufficient to achieve good performance.

Once the regularizer is learned, its parameters $\phi_{\text{META}}$ are frozen and the final student network initialized from scratch is trained on the target dataset using the regularized loss function $F$. This meta-testing procedure generalizes the learned knowledge to the target dataset with $R$ parameterized by $\phi_{\text{META}}$ as summarized in Algorithm 2.

Our meta-learning approach is general and can be implemented by any type of regularizer. In this paper, we use
where the marginal log-likelihood can be lower bounded:

\[
\begin{align*}
\mathbb{E} &\geq \frac{1}{N} \sum_{n=1}^{N} \log P(x_n^S | \theta) = \sum_{n=1}^{N} \log P(x_n^S | \theta_n) - KL(q(\theta_n | D_S) \| P(\theta | \phi)). 
\end{align*}
\]

Proposition 1 indicates its lower bound.

Proposition 2. The parameters \( \phi \) can be estimated by maximizing the evidence lower-bound of \( \log P(D_S | \phi) \) via expectation maximization (EM) as shown in Algorithm 1.

\[\begin{align*}
\hat{\theta}_k &\approx \arg\max_{\theta} \mathcal{E}(q_k | \theta_k), \\
\phi_{k+1} &\approx \arg\max_{\phi} \mathcal{E}(q_k | \phi_k).
\end{align*}\]

Algorithm 1 Meta-training for learning priors.
\begin{itemize}
  \item \textbf{Input:} Batch size \( N \), \# of iterations \( K \), learning rate \( \alpha \).
  \item \textbf{Input:} \# of inner iterations \( l \), meta learning rate \( \beta \).
  \item Initialize \( \theta_0, \phi_0 \).
  \item for \( k = 0 \) to \( K-1 \) do
  \item Sample \( N \) examples \( \{(x_n^S, \tilde{x}_n^S, y_n^S) \sim D_S\}^{N}_{n=1} \).
  \item \( \hat{\theta}_k \leftarrow \theta_k \).
  \item for \( l = 0 \) to \( l-1 \) do
  \item \( \theta_{l+1} \leftarrow \theta_l - \alpha \nabla_{\theta_l} \mathcal{L}(x_n^S, \tilde{x}_n^S, y_n^S; \theta_k) \) \text{ E-step}
  \item \( \phi_{k+1} \leftarrow \phi_k - \beta \nabla_{\phi_k} \mathcal{E}(x_n^S, \tilde{x}_n^S, y_n^S; \theta_k) \) \text{ M-step}
  \item \( \theta_{k+1} \leftarrow \theta_k - \alpha \nabla_{\theta_k} \mathcal{L}(x_n^S, \tilde{x}_n^S, y_n^S; \theta_k) \)
  \item end for
  \item end for
  \item \( \phi_{\text{META}} \leftarrow \phi_K \).
\end{itemize}

Algorithm 2 Meta-testing for knowledge generalization.
\begin{itemize}
  \item \textbf{Input:} Batch size \( N \), \# of iterations \( K \), learning rate \( \alpha \).
  \item \textbf{Input:} Learned parameters \( \phi_{\text{META}} \) from Algorithm 1.
  \item Initialize \( \theta_0 \).
  \item for \( k = 0 \) to \( K-1 \) do
  \item Sample \( N \) examples \( \{(x_n^T, y_n^T) \sim D_T\}^{N}_{n=1} \).
  \item \( \theta_{k+1} \leftarrow \theta_k - \alpha \nabla_{\theta_k} \mathcal{L}(x_n^T, y_n^T; \theta_k, \phi_{\text{META}}) \)
  \item end for
  \item \( \theta_{\text{DST}} \leftarrow \theta_K \).
\end{itemize}

Both E-step and M-step are solved by gradient descent as commented in Algorithm 1. We have thus shown that Algorithm 1 is an instance of EM.

5. Experiments

The proposed approach is evaluated in 3D hand pose estimation. We aim to answer the following three questions: (1) Can our Cross-Modal Knowledge Distillation (CMKD) distill accurate cross-modal knowledge from the source dataset? (Sect. 5.3) (2) Does the proposed Cross-Modal Knowledge Generalization (CMKG) successfully transfer learned knowledge to the target dataset? (Sect. 5.4) (3) And what factors influence the effect of our CMKG? (Sect. 5.5)

5.1. Implementation Details

For simplicity, we use the same architecture for teacher and student networks. We choose ResNet [15] as the backbone, and adjust the final fully connected layer to output a vector representing the 3D positions of 21 hand joints. All corresponding depth maps of RGB images in the dataset are employed as the modality containing superior knowledge.

Data Augmentation. Recent methods [4, 12, 46] show that learning from synthetic data improves the performance
of 3D pose estimation, as it offers more effective hand variations than traditional data augmentation technologies, e.g., random cropping and rotation. Hence, we create a synthetic dataset of paired hand images and depth maps with their 3D annotations using the MANO [37] hand model for synthetic data augmentation. Following the setting of [4], hand geometries are obtained by sampling pose and shape parameters from $[-2, 2]^{10}$ and $[-0.03, 0.03]^{10}$, respectively. Meanwhile, hand appearances are modeled by the original scans with 3D coordinates and RGB values from [37]. We create example hand appearances using these registered scan topologies. After rotations, translations and scalings are applied to hand models, the textured hands are finally rendered on background images which are randomly sampled and cropped from [18, 24]. In total, we synthesize 50,000 hand images with large variations for training.

Network Training. The input image is resized to $256 \times 256$. For CMKD, all networks are trained using Adam [19] with mini-batches of size 32. The learning rate is set as $2.5 \times 10^{-4}$. The teacher is pre-trained for 200 epochs, while the student is trained with only the regression loss for 100 epochs and then fine-tuned with the full loss for another 100 epochs. For CMKG, the regularizer is optimized using Stochastic Gradient Descent (SGD) with the learning rate of $1.0 \times 10^{-3}$ during the fine-tuning of the student network.

5.2. Datasets and Metrics

Our proposed approach is comprehensively evaluated on two publicly available datasets for 3D hand pose estimation: RHD [59] and STB [55] with the standard metrics.

Datasets. Rendered Hand Pose Dataset (RHD) [59] is a synthetic dataset built upon 20 different characters performing 39 actions. It provides 41,258 images for training and 2,728 images for evaluation with a resolution of $320 \times 320$. All of them are fully annotated with a 21 joint skeleton hand model and additionally the depth map for each hand. This dataset is challenging due to the large variations in viewpoints and textures. We employ RHD for training and evaluating our knowledge distillation method.

Stereo Hand Pose Tracking Benchmark (STB) [55] is a real-world dataset which contains 18,000 stereo image pairs as well as the ground truth 3D positions of 21 hand joints from different scenarios. This benchmark has 12 different sequences and every sequence contains 1,500 stereo pairs. Following the evaluation protocol of [5, 12, 40, 59], we use the sequence of B1 for evaluation and the others for training. STB is utilized for evaluating the proposed cross-modal knowledge generation algorithm.

To make the joint definition consistent across different datasets, we reorganize the joints of each hand according to the layout of MANO [37]. Especially, we move the root joint location from palm center to wrist of each hand in STB. Following the same protocol used in [5, 12, 40, 59], the absolute depth of root joint (wrist) and global hand scale, which is set as the bone length between MCP and PIP joints of the middle finger, are provided at test time.

Metrics. We evaluate the performance of 3D hand pose estimation with three common metrics in the literature: (1) EPE: the mean hand joint error which measures the average Euclidean distance in millimeters (mm) between the predicted 3D joints and the ground truth; (2) 3D PCK: the percentage of correct key points which are within the Euclidean distance of a certain threshold to its respective ground truth position; (3) AUC: the area under the curve on 3D PCK for different error thresholds.

5.3. Evaluation of Knowledge Distillation

To evaluate the performance of the proposed knowledge distillation approach for 3D hand pose estimation, we train three networks for each setting: a baseline network trained with RGB images (RGB), a teacher network trained with depth maps (Depth) and a student network trained using the knowledge distillation algorithm presented in Sect. 3 (KD). All the experiments are conducted on RHD dataset.

Ablation Study. We first evaluate the impacts of different losses used in knowledge distillation, data augmentation, and network architecture on the performance of 3D hand pose estimation. The results of EPE are presented in Table 1. We can see that the model trained with the full distillation loss ($L_{ACT}$ and $L_{ATT}$) achieves higher performance improvement, from 1.27 (mm) to 2.49 (mm), which indicates that all the losses have contributions to distilling cross-modal knowledge from depth maps for 3D hand pose estimation. Moreover, synthetic data augmentation and employing deeper network during the training procedure can further boost the performance.

Comparison to State of the Art. We compare the 3D PCK curves with state-of-the-art methods [5, 40, 52, 53, 59] on RHD dataset in Fig. 4. We use ResNet-50 as the backbone. Note that some other works [4, 12, 56] aim to predict the 3D hand shape other than hand joint locations, which are with different research targets compared with ours. Therefore, they are not included here. In Fig. 4 (left), our method surpasses most existing methods except [5], which has a higher AUC of 0.015. However, it is not directly compa-

<table>
<thead>
<tr>
<th>Settings</th>
<th>Backbone</th>
<th>EPE (RGB / Depth / KD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_{ACT}$</td>
<td>ResNet-18</td>
<td>24.68 / 13.60 / 23.41</td>
</tr>
<tr>
<td>$L_{ACT}$, $L_{ATT}$</td>
<td>ResNet-18</td>
<td>24.68 / 13.60 / 22.19</td>
</tr>
<tr>
<td>$L_{ACT}$, $L_{ATT}$, $A$</td>
<td>ResNet-18</td>
<td>23.07 / 12.06 / 20.89</td>
</tr>
<tr>
<td>$L_{ACT}$, $L_{ATT}$, $A$</td>
<td>ResNet-50</td>
<td>20.74 / 10.78 / 18.00</td>
</tr>
</tbody>
</table>
Figure 4. Comparisons with state of the art. Left: 3D PCK on RHD [59] of our knowledge distillation approach (CMKD). Our method has comparable performance to Cai et al. [5] which relies on additional 2D annotations for network training. Middle: Comparison with Yuan et al. [53] which also distills knowledge from depth. Our approach obtains a more significant improvement (red area, $\Delta_{\text{AUC}} = 0.045$) than [53] (green area, $\Delta_{\text{AUC}} = 0.041$). Right: Our knowledge generalization method (CMKG) obtains state-of-the-art results on STB [55].

Table 2. Effect of different classes of regularization functions on STB [55]. Note that $\sigma$ denotes the constant weight manually chosen for the default $\ell^1$ or $\ell^2$ regularizer. We report EPE (mm) and AUC together with the performance gain for each method.

<table>
<thead>
<tr>
<th>Regularizer</th>
<th>EPE (mm)</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>15.67</td>
<td>0.915</td>
</tr>
<tr>
<td>$\ell^1$, $\sigma = 1.0 \times 10^{-3}$</td>
<td>11.41</td>
<td>0.972</td>
</tr>
<tr>
<td>$\ell^2$, $\sigma = 1.0 \times 10^{-6}$</td>
<td>11.82</td>
<td>0.964</td>
</tr>
<tr>
<td>$\ell^2$, $\sigma = 1.0 \times 10^{-3}$</td>
<td>12.28</td>
<td>0.957</td>
</tr>
<tr>
<td>$\ell^2$, $\sigma = 1.0 \times 10^{-5}$</td>
<td>12.02</td>
<td>0.964</td>
</tr>
</tbody>
</table>

$\mathcal{R}$, $\ell^1$-regularized

$\mathcal{R}$, $\ell^2$-regularized

Table 5. Histograms of the parameters learned by different regularization functions on STB [55] dataset. Left: Histogram of the network $f$ without any form of regularization. Right: Histogram of the network trained with the proposed regularizer $\mathcal{R}$ in Eq. (10).
the proposed CMKG to other approaches [27, 31, 40, 59] on STB in Fig. 4 (right). We can see that our regularized network matches the state-of-the-art performance without using complex network architecture, loss functions or additional constraints like previous methods. Our visual results are shown in Fig. 6. As seen, our method is able to accurately predict 3D hand poses across different datasets and generalize the learned knowledge to some novel cases.

5.5. Discussion

One potential concern about the learned knowledge (regularizer) from the source dataset is how it performs when applied to different target datasets. First of all, we point out that it is impossible to learn a domain-independent regularizer from a single source which performs consistently well on all other datasets, since their data usually follow different statistics. Here, we hypothesize that the effect of the learned regularizer depends on two factors: (1) the domain shift between the source and target dataset, and (2) the effect of regularization on the target dataset.

The first factor is straightforward as large domain shifts always lead to difficulties in network generalization. This is a well-defined problem in transfer learning which is tackled by domain adaptation [30, 33, 34]. To illustrate the second factor, we conduct an additional experiment which applies the same regularizers in Sect. 5.4 to a number of different target datasets. Due to the space limitation, we ask the readers to refer to the supplementary material for detailed setups of this experiment. Looking at Table 3, we see a strong correlation between the default and the proposed regularizer: if there is a large increase obtained by the default regularizer, \( R \) can boost the performance even further; otherwise, our improvement is limited. This is intuitive since our formulation is consist with the default regularization technique.

![Figure 6. Visual results of our approach on RHD [59] (top) and STB [55] (middle). To demonstrate the generalizability of the proposed method, we also show the results after applying the network trained on STB to the synthetic dataset (bottom). Best viewed in color.](image)

<table>
<thead>
<tr>
<th>Target Dataset</th>
<th>w/o ( R )</th>
<th>Default ( \ell^2 )</th>
<th>( R ) in Eq. (10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FreiHAND [60], ( G )</td>
<td>12.37</td>
<td>12.28,0.09</td>
<td>12.27,0.19</td>
</tr>
<tr>
<td>FreiHAND [60], ( H )</td>
<td>14.49</td>
<td>14.02,0.47</td>
<td>13.82,0.67</td>
</tr>
<tr>
<td>FreiHAND [60], ( S )</td>
<td>15.80</td>
<td>14.92,0.88</td>
<td>14.26,1.54</td>
</tr>
<tr>
<td>FreiHAND [60], ( A )</td>
<td>16.18</td>
<td>15.16,1.02</td>
<td>14.18,2.00</td>
</tr>
<tr>
<td>STB [55]</td>
<td>15.67</td>
<td>12.02,1.65</td>
<td>8.18,7.49</td>
</tr>
</tbody>
</table>

Table 3. Effect of regularizers on different target datasets. We report EPE (mm) and the performance gain for each setting. \( G \), \( H \), \( S \) and \( A \) are four different domains contained in FreiHAND [60].

Our findings suggest multiple directions of future work. For one, the proposed scheme currently has access to only one single source dataset; we believe that learning from multiple sources will result in better generalizability of the model. On the other hand, we treat target priors as a regularization term in this work, which is perhaps the simplest formulation. We believe that a further exploration on choices of this term will result in improved performance.

6. Conclusion

We introduce an end-to-end scheme for Cross-Modal Knowledge Generalization to transfer cross-modal knowledge between source and target datasets where superior modalities are missing. The core idea is to interpret knowledge as priors on the parameters of the student network which can be efficiently learned by meta-learning. Our method is comprehensively evaluated in 3D hand pose estimation. We show that our scheme can efficiently generalize cross-modal knowledge to the target dataset and significantly boost the network to match the state-of-the-art performance. We believe our work provides new insights in conventional cross-modal knowledge distillation tasks, and serves as a strong baseline in this novel research direction.
References


