Probabilistic Model Distillation for Semantic Correspondence

Xin Li[1] Deng-Ping Fan[2] Fan Yang[1,*] Ao Luo[3] Hong Cheng[4] Zicheng Liu[5]
[1] Group 42 (G42) [2] Inception Institute of AI [3] Megvii Technology [4] UESTC [5] Microsoft

Abstract

Semantic correspondence is a fundamental problem in computer vision, which aims at establishing dense correspondences across images depicting different instances under the same category. This task is challenging due to large intra-class variations and a severe lack of ground truth. A popular solution is to learn correspondences from synthetic data. However, because of the limited intra-class appearance and background variations within synthetically generated training data, the model's capability for handling "real" image pairs using such strategy is intrinsically constrained. We address this problem with the use of a novel Probabilistic Model Distillation (PMD) approach which transfers knowledge learned by a probabilistic teacher model on synthetic data to a static student model with the use of unlabeled real image pairs. A probabilistic supervision reweighting (PSR) module together with a confidence-aware loss (CAL) is used to mine the useful knowledge and alleviate the impact of errors. Experimental results on a variety of benchmarks show that our PMD achieves state-of-the-art performance. To demonstrate the generalizability of our approach, we extend PMD to incorporate stronger supervision for better accuracy – the probabilistic teacher is trained with stronger key-point supervision. Again, we observe the superiority of our PMD. The extensive experiments verify that PMD is able to infer more reliable supervision signals from the probabilistic teacher for representation learning and largely alleviate the influence of errors in pseudo labels. Code is available at https://github.com/fanyang587/PMD.

1. Introduction

Matching all pixels between images is a classic research problem in computer vision. Unlike stereo matching [62] or optical flow [3] that deal with images containing different viewpoints of one scene or object, semantic correspondence poses additional challenges by pushing the boundaries of dense matching to correspondence estimation between visually similar images. Matching beyond scene leads to many meaningful applications, such as attribute transfer [34, 44], image editing [2, 17, 40], scene collaging [27], object discovery and segmentation [43, 59]. This task is extremely challenging due to large intra-class variation and a severe lack of groundtruth correspondence maps.

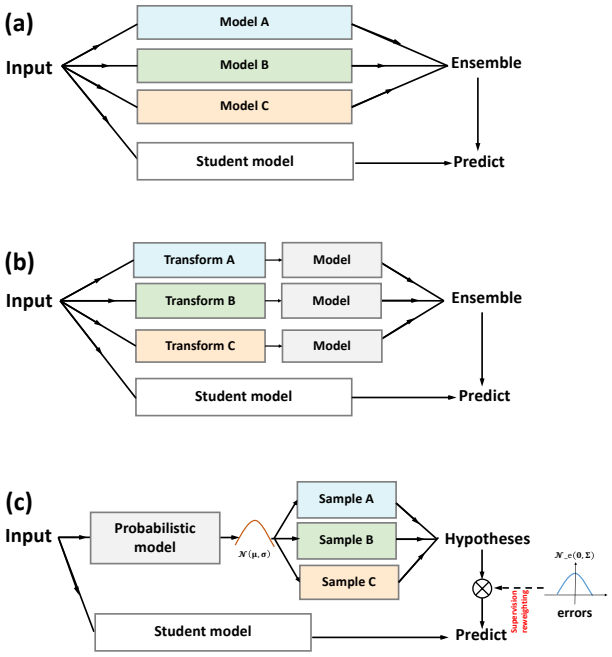

Figure 1. **Idea Illustration**. Instead of **(a)** ensembling multiple models to generate "soft" predictions as the student's target [22], or **(b)** creating a single "hard" label from different transformations [54], **(c)** our idea is to distill knowledge from hypotheses from a probabilistic teacher model in a probabilistic manner.

One way to overcome these challenges is to use hand-crafted image descriptors (*e.g.*, SIFT [50], HOG [9], DAISY [67]) in combination with different regularization methods [18, 31, 45–47, 53, 61, 66] to estimate semantic correspondence. Unfortunately, hand-crafted descriptors are intrinsically weak in capturing high-level semantics and thus, less robust to large intra-class variation, geometric deformation and background clutter. Inspired by the success of self-supervised deep learning models, a series of approaches have been proposed to learn semantic descriptors [63], or directly regress parameters of a global trans-

*Corresponding author: Fan Yang *(fanyang_uestc@hotmail.com)*

formation model [55, 64], by using synthetically generated data. Although these approaches achieve state-of-the-art performances, using synthetic data could weaken the model's ability to deal with complex intra-class and background variations for real pairs due to limited appearance variation in synthetic training pairs. To alleviate this problem, recent approaches use different types of auxiliary annotations (*e.g.*, key points [5, 6, 8, 29, 37, 42], semantic masks [7, 39], 3D CAD model [68, 70], and image-level labels [26, 32, 56, 57]) as supervision signals for model training on real pairs. However, auxiliary annotations not only are labor-intensive to collect, but also could weaken the generalization ability of the learned network models. The observations above inspire us to wonder: *Can the knowledge learned from synthetic data be generalized and even enhanced to deal with real image pairs without using any manual annotation?*

To answer this question, as shown in Figure 1, we propose to perform knowledge distillation within a probabilistic teacher-student (PTS) framework. Our idea is to generate pseudo semantic flows on *unlabeled real pairs* using a teacher model trained on *synthetic data*, and then train the student model using the pseudo-groundtruth flows. However, training the student model on the predictions by a single static teacher model would inevitably encode noisy, biased information, which causes the student to be upper-bounded by its teacher model. To address this problem, we employ a probabilistic teacher model to provide multiple/diverse matching hypotheses for each real pair, and use a student model working with a probabilistic supervision reweighting (PSR) module to distill 'correct' (pixel-wise) supervision signals from pseudo correspondence maps. Moreover, a confidence-aware loss (CAL) is introduced to further reduce the impact of errors on pseudo flows during knowledge distillation.

Our approach, called Probabilistic Model Distillation (PMD), can successfully distill the knowledge from hypotheses of the probabilistic teacher model trained on *synthetic data*, and safely generalize it into a powerful student model with potentially unlimited real-world image pairs in an *unsupervised* manner. We demonstrate that the model trained with our probabilistic model distillation produces better results than all state-of-the-art (SOTA) self-supervised approaches, even those utilizing (auxiliary) manual annotations.

To further demonstrate the advantage and generalizability of our approach, we extend it to incorporate stronger supervision – the probabilistic teacher is trained using the (auxiliary) matched key points. Surprisingly, even with the same training set, the student model outperforms the teacher by a large margin and sets new records on multiple benchmarks. The contributions of this work are summarized as:

- **A novel probabilistic approach for model distillation.** To our best knowledge, this is the first attempt to distill knowledge from multiple/diverse hypotheses produced by a probabilistic teacher model for self-supervised/unsupervised model learning. This approach is able to generalize the knowledge learned from *synthetic data* into a new model for better handling *real-world data*.
- **A new probabilistic teacher-student network for semantic correspondence.** We present the probabilistic teacher-student network, an effective instantiation of our probabilistic model distillation for semantic correspondences, which consists of a probabilistic teacher model learning knowledge from *synthetic data*, and a static student model working with a novel probabilistic supervision reweighting (PSR) module to perform distillation from multiple hypotheses of each *real image pair*.
- **State-of-the-art results on widely-used benchmarks.** The model trained with our probabilistic model distillation outperforms state-of-the-art methods on a variety of benchmarks, *e.g., PF-WILLOW, PF-PASCAL* and *SPair-71k* . Our method even surpasses approaches that require extra (auxiliary) annotations for model learning.
- **Potential generalizability to stronger supervision**. We show that our approach can be extended to incorporate stronger supervision signals for better accuracy. With the strongly supervised teacher model, the student model can achieve better performance, outperforming existing state-of-the-art methods with different degrees of supervision.

2. Related Work

Knowledge Distillation. Knowledge distillation is mostly used to learn small and compact models [4, 22, 58, 69] in various practical applications. Romero *et al*. [58] introduce FitNet to compress wide and deep networks into thinner and deeper ones. Zagoruyko *et.al* [69] propose attention as a mechanism of transferring knowledge from the teacher model to the student model. In [4], an end-to-end trainable framework is introduced to compact multi-class object detection models. Radosavovic et.al [54] use the data distillation technique to exploit omni-supervised setting for multiple tasks. In addition to learning compact models, distillation technique has been explored for transferring knowledge between different domains [12, 16], or from an ensemble of models to a single student model [36, 41]. Unlike prior works, our novelty lies in using diverse hypotheses of a single probabilistic model trained on *synthetic data* as supervisions to learn a static model on *unlabeled real data*.

Dense Semantic Correspondence. The goal of semantic correspondence is to build correspondences between semantically similar images. Liu *et al.* [47] pioneer the idea of *densely matching across semantically similar scenes*, and present SIFT flow [45]. This approach has been further improved with more carefully designed descriptors [20, 67] or

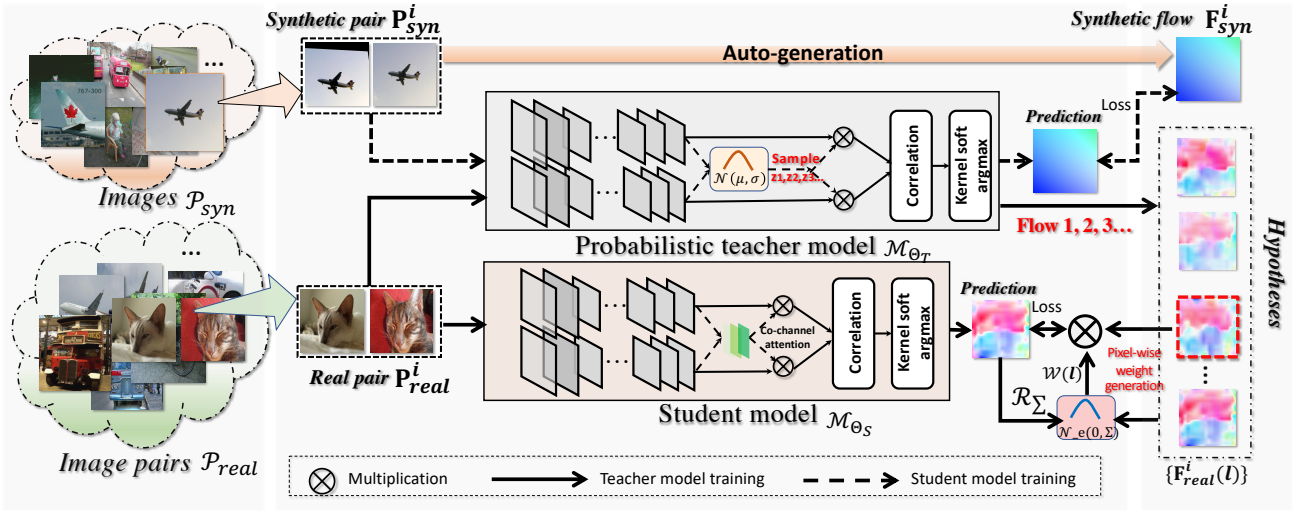

Figure 2. **Main Pipeline**. Our probabilistic teacher-student network consists of a probabilistic teacher model, a static student model and a probabilistic supervision reweighting module. The probabilistic teacher model is trained on *synthetic pairs* while the student model is trained on *real pairs* with our confidence-aware loss (CAL). Please refer to § 3 for more details.

graphical models [18, 31, 46, 53, 61, 66]. However, with hand-crafted features, these approaches often suffer from large intra-class variation, geometric deformation and background clutter for the lack of semantics in feature representations. Recent works overcome this issue by utilizing deep neural networks to extract semantic features [49, 63] or regress parameters of a global transformation model [55, 64]. Due to the severe lack of groundtruth correspondence maps, existing models are trained with either synthetic data in a self-supervised manner [49, 55, 63, 64] or real data with auxiliary annotations (*e.g.,* key points [5, 6, 8, 29, 37, 42], semantic masks [7, 39], 3D CAD model [68, 70], image-level labels [26, 32, 56, 57]). In this paper, we show that the knowledge for semantic correspondence acquired from the *synthetic* data can be generalized to the distilled model with *unlabeled real* data to better handle difficult real-world challenges. Moreover, the strong supervision signals can be greatly enhanced with our approach to better guide the representation learning of the student model.

3. Methodology

3.1. Preliminaries

Problem Definition. Let the semantic matching model be represented by the function $\mathcal{M}_\Theta$ parameterized by weights Θ, that takes an image pair $\mathbf{P}^i_{real} = (\mathbf{I}^s_i, \mathbf{I}^t_i) \in \mathcal{P}_{real}$ as input, and generates the semantic flow $\mathbf{F}^i_{real} = \mathcal{M}_\Theta(\mathbf{P}^i_{real}; \Theta) \in \mathcal{F}_{real}$ which reflects the pixel-wise correspondences between source image $\mathbf{I}^s_i$ and target image $\mathbf{I}^t_i$. Our goal is to learn Θ for achieving the accurate semantic flow when given only *unlabeled* image pairs.

Overview. To achieve this goal, we propose to perform *knowledge distillation* within a probabilistic teacher-student (PTS) framework, as shown in Figure 2. The probabilistic teacher network $\mathcal{M}_{\Theta_T}$, parametrized by weights Θ_T, is first trained on a data set with *synthetic* image pairs $\mathcal{P}_{syn}$ and groundtruth flows hypotheses $\mathcal{F}_{syn}$ generated synthetically using random transformations of the same image. Then, multiple/diverse pseudo flows $\{\mathbf{F}^i_{real}(l)\}^m_{l=1}$ are generated by the trained teacher $\mathcal{M}_{\Theta_T}$ for each real image pair $\mathbf{P}^i_{real} \in \mathcal{P}_{real}$. To optimize Θ_S, the static student network $\mathcal{M}_{\Theta_S}$ works with a probabilistic supervision reweighting (PSR) module $\mathcal{R}_\Sigma$ to gradually infer the more "correct" supervision signal at each position among $\{\mathbf{F}^i_{real}(l)\}^m_{l=1}$ for parameter learning, under the supervision of a confidence-aware loss (CAL).

3.2. Probabilistic Teacher-Student Network

Semantic Matching Network. We start by introducing our basic matching model for semantic correspondence. Similar to most deep matching models [39, 55, 63, 64], our model employs a siamese network architecture, which takes a pair of images as input and generates the pixel-wise flow map. As shown in Figure 3 (a), we follow [39, 55, 57] to use a ResNet-101 [21] pretrained on ImageNet [11] with two additional convolution layers to extract features from image pairs $\mathbf{P} = (\mathbf{I}^s, \mathbf{I}^t)$: $\mathbf{P} \mapsto (\mathbf{d}^s, \mathbf{d}^t)$. Inspired by [24], a lightweight co-channel attention module, consisting of a concatenation layer, two 1×1 convolution layers, a global average pooling layer and a sigmoid activation function, is used to enhance the representational power of the extracted features, *i.e.,* $(\mathbf{d}^s, \mathbf{d}^t) \mapsto (\mathbf{D}^s, \mathbf{D}^t)$. Finally, based on the enhanced features, one correlation layer [14] followed by the kernel soft argmax [39] is employed to generate the semantic flow, *i.e.,* $(\mathbf{D}^s, \mathbf{D}^t) \mapsto \mathbf{F}$. The training of our matching model can be formulated as a regression problem,

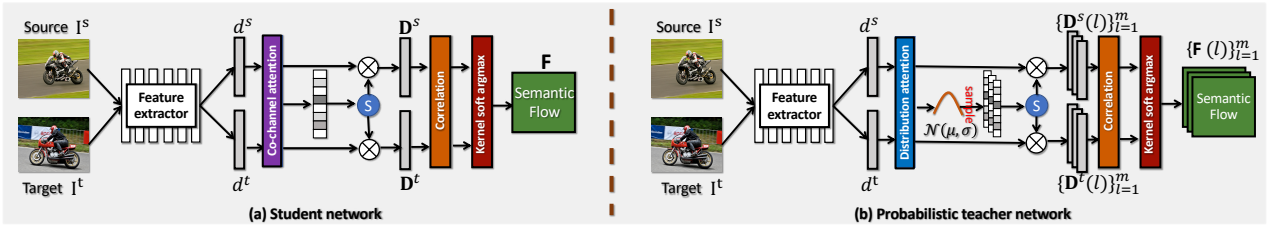

Figure 3. **Network architectures**. **(a)** Semantic matching network (student model). **(b)** Probabilistic matching network (teacher model).

which, theoretically, needs "groundtruth" semantic flows as supervision signals. Since it is difficult to acquire the groundtruths for real image pairs, we provide a reasonable alternative for model training by using a novel distillation technique within a probabilistic teacher-student framework.

Probabilistic Teacher Network. To overcome the severe lack of groundtruth correspondence maps on *real pairs*, we use a teacher model trained on *synthetic data* to provide the supervision signals. However, such a teacher model usually provides unreliable and biased predictions for *real pairs*, which would cause the student model to be misguided and upper-bounded. One way to alleviate this problem is to ensemble different teacher models to provide diverse and individually accurate predictions as in [1, 13, 23, 65], which, unfortunately, is time-consuming for training. Unlike these approaches, we draw inspiration from the variational bayesian approach [35] and learn a single probabilistic teacher network to mimic a group of models.

Our probabilistic teacher network $\mathcal{M}_{\Theta_T}$ basically shares the same architecture as our semantic matching network. The difference is that what our probabilistic teacher model delivers to each position is no longer a single flow vector but multiple/diverse hypotheses. Since the flow is intrinsically generated by feature matching, we propose to model attention distributions in the extracted feature maps as Gaussian, parameterized by mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}$. Intuitively, enriching the extracted features with different co-attention information could enable dense correspondence fields to exhibit different tendencies, so that meaningful hypotheses can be produced. Therefore, as shown Figure 3 (b), we replace the co-channel attention module with a *distribution attention module* which has two additional branches to learn the distributions. From the probabilistic perspective, each attention feature is treated as a random variable. Mathematically, the distribution attention module $\mathcal{A}_\Upsilon$, parametrized by weights Υ, learns the probability of these variants for the extracted features $(\mathbf{d}^s, \mathbf{d}^t)$. This distribution, denoted as $\boldsymbol{Z}$, is modelled as a Gaussian with mean $\boldsymbol{\mu}\big((\mathbf{d}^s, \mathbf{d}^t); \Upsilon\big) \in \mathbb{R}^c$ and variance $\boldsymbol{\sigma}\big((\mathbf{d}^s, \mathbf{d}^t); \Upsilon\big) \in \mathbb{R}^c$. To generate a set of attention features, we apply the teacher network to the same input pair for m times . At each iteration $l \in \{1, 2, \ldots, m\}$, we get a random sample $\boldsymbol{z}_l \in \mathbb{R}^c$ from distribution $\boldsymbol{Z}$,

$$\begin{aligned} \boldsymbol{z}_l &\sim \boldsymbol{Z}(\cdot \mid (\mathbf{d}^s, \mathbf{d}^t)) \\ &= \mathcal{N}\Big(\boldsymbol{\mu}\big((\mathbf{d}^s, \mathbf{d}^t); \Upsilon\big), \mathrm{diag}\big(\boldsymbol{\sigma}((\mathbf{d}^s, \mathbf{d}^t); \Upsilon)\big)\Big). \end{aligned} \tag{1}$$

The sample $\boldsymbol{z}_l$ is then used to enhance the extracted features $(\mathbf{d}^s, \mathbf{d}^t)$, *i.e.,* $\mathbf{D}^s(l) = \boldsymbol{f}_e(\mathbf{d}^s, \boldsymbol{z}_l)$ and $\mathbf{D}^t(l) = \boldsymbol{f}_e(\mathbf{d}^t, \boldsymbol{z}_l)$ where $\boldsymbol{f}_e$ means the channel-wise multiplication operation, similar to [24]. Finally, a function $\boldsymbol{f}_m$ composed of a correlation layer [14] and a kernel soft argmax [39] takes $(\mathbf{D}^s(l), \mathbf{D}^t(l))$ and maps them to the semantic flow $\mathbf{F}(l)$.

(1) Training with Synthetic Data. Since the groud-truth correspondence maps for real data are unavailable, we are motivated by self-supervised learning approaches [55, 63, 64] and create a synthetic training `train` set $\mathcal{P}_{syn}$ including N_{syn} synthetic image pairs $\{\mathbf{P}^i_{syn}\}_{i=1}^{N_{syn}} \in \mathcal{P}_{syn}$ and correspondence maps $\{\mathbf{F}^i_{syn}\}_{i=1}^{N_{syn}} \in \mathcal{F}_{syn}$ (by applying random affine transformations on each single image). Therefore, our probabilistic teacher network $\mathcal{M}_{\Theta_T}$ can be trained with synthetic data using the following loss function:

$$\begin{aligned} \mathcal{L}_T(\Theta_T, \Upsilon) = \mathcal{L}_{\mathbf{mse}}(\mathbf{F}_{syn}, \mathbf{F}(l)) + \alpha \cdot \mathcal{L}_{\mathbf{smooth}}(\mathbf{F}(l)) \\ + \beta \cdot \mathcal{D}_{\mathbf{KL}}(\boldsymbol{Z}(\cdot \mid (\mathbf{d}^s, \mathbf{d}^t)) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I})), \end{aligned} \tag{2}$$

where the loss for probabilistic teacher network $\mathcal{L}_T$ is a weighted combination of a standard mean squared error (MSE) loss $\mathcal{L}_{\mathbf{mse}}$, a smoothness loss [39] $\mathcal{L}_{\mathbf{smooth}}$ and a Kullback-Leibler (KL) divergence $\mathcal{D}_{\mathbf{KL}}$.

(2) Hypotheses Generation with Real Data. Given N_{real} *unlabeled real pairs* $\mathcal{P}_{real}$, we can use the trained probabilistic teacher network to generate m diverse pseudo-groundtruth flows (hypotheses) $\{\check{\mathbf{F}}^i_{real}(l)\}_{l=1}^m \in \check{\mathcal{F}}_{real}$ for each of them $\mathbf{P}^i_{real} \in \mathcal{P}_{real}$. Our goal is to distill knowledge from these pseudo flows to train a powerful student model.

Student Network. We take the semantic matching network as our student model, denoted as $\mathcal{M}_{\Theta_S}$, and use the pseudo flows $\check{\mathcal{F}}_{real}$ on the real-world pairs $\mathcal{P}_{real}$ by $\mathcal{M}_{\Theta_T}$ to learn its parameters Θ_S. Instead of "naïvely" fusing them by average as supervision signals, we use a novel probabilistic model distillation (PMD) approach to gradually infer more "correct" (pixel-wise) supervision signals from $\{\check{\mathbf{F}}^i_{real}(l)\}_{l=1}^m \in \check{\mathcal{F}}_{real}$. Next, we detail our PMD approach.

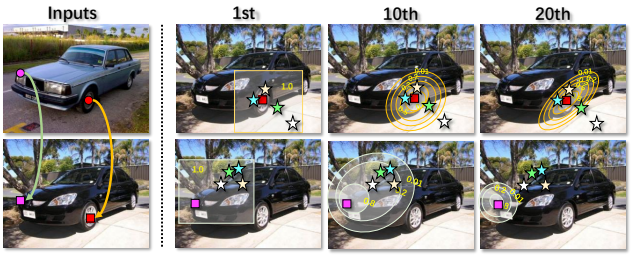

Figure 4. **Illustration of probabilistic supervision reweighting**. For each position, a learnable error distribution is used to guide the supervision reweighting of all hypothese ('★'). The more 'correct' signals are expected to have higher weights while the incorrect ones have lower weights.

3.3. Probabilistic Model Distillation

For each real-world pair $\mathbf{P}^i_{real} \in \mathcal{P}_{real}$, the probabilistic teacher model produces m pseudo-groundtruth flows $\{\check{\mathbf{F}}^i_{real}(l)\}^m_{l=1} \in \check{\mathcal{F}}_{real}$ as candidate labels for training the student model. Instead of naïvely ensembling all pseudo flows (*e.g.,* by average) as the student's target, our idea is to automatically measure the correctness of *pixel-wise* pseudo labels, and assign higher weights to more correct ones with a learnable probabilistic supervision reweighting (PSR) module.

Probabilistic Supervision Reweighting. Inspired by [30], we formulate the supervision reweighting for pseudo-groundtruth flows as a *learning-with-multiple-labels* problem and infer the correctness by modeling the *pixel-wise* error distributions within them, as shown in Figure 4.

(1) Probabilistic Error Modeling. Given a real image pair $\mathbf{P}^i_{real} \in \mathcal{P}_{real}$ and a corresponding pseudo flow $\check{\mathbf{F}}^i_{real}(l)(p) = (\check{u}^i_{real}(l)(p), \check{v}^i_{real}(l)(p)) \in \check{\mathcal{F}}_{real}$, the error at position p can be represented as: $e^i(l)(p) = \mathbf{F}^i_{real}(p) - \check{\mathbf{F}}^i_{real}(l)(p)$, where $\mathbf{F}^i_{real}(p) = (u^i_{real}(p), v^i_{real}(p)) \in \mathcal{F}_{real}$ denotes the correct flow. We assume that the errors at each position follow the zero-mean Gaussian distribution $\boldsymbol{\xi}^i_p$: $e^i(l)(p) \sim \boldsymbol{\xi}^i_p$, and only depend on the image pair $\mathbf{P}^i_{real}$. Furthermore, as the flow field is a two-dimensional vector, we treat errors in each dimension independently. Therefore, there are a total of $h \times w \times 2$ distributions for each flow map, modeled by $\Sigma^i = \{\sigma^i_u(p), \sigma^i_v(p)\}^{h\times w}_{p=1}$. The probabilistic supervision reweighting (PSR) module $\mathcal{R}_{\boldsymbol{\Sigma}}$ therefore includes $\boldsymbol{\Sigma} = \{\Sigma^i\}^{N_{real}}_{i=1}$ learnable parameters, which are used to parameterize the pixel-wise error distributions of all (N_{real}) real pairs. Assuming error is known, then we can estimate the weight of pseudo-groundtruth flow $\check{\mathbf{F}}^i_{real}(l)$ at position p by the following equations:

$$\begin{cases} \mathcal{W}^i_u(l)(p) = e^{-\frac{||e^i(l_u)(p)||^2}{(\sigma^i_u(p))^2}} = e^{-\frac{||u^i_{real}(p) - \check{u}^i_{real}(l)(p)||^2}{(\sigma^i_u(p))^2}}, \\ \mathcal{W}^i_v(l)(p) = e^{-\frac{||e^i(l_v)(p)||^2}{(\sigma^i_v(p))^2}} = e^{-\frac{||v^i_{real}(p) - \check{v}^i_{real}(l)(p)||^2}{(\sigma^i_v(p))^2}}, \end{cases} \tag{3}$$

where $\mathcal{W}^i_u(l)(p)$ and $\mathcal{W}^i_v(l)(p)$ mean the weight for $\check{\mathbf{F}}^i_{real}(l)$ of each dimension at position p.

(2) Joint Model Distillation and Supervision Reweighting. We connect our PSR module $\mathcal{R}_{\boldsymbol{\Sigma}}$ with the student network $\mathcal{M}_{\Theta_S}$ to learn both the error parameters $\boldsymbol{\Sigma}$ and the student model parameters Θ_S. Recall that our goal is to assign higher weights to the more correct labels (at pixel-wise level) for training, on the condition that groud truths are unknown. It turns out that the expectation-maximization (**EM**) algorithm [10] can be applied to accomplish our goal, resulting in a procedure which iterates between error model parameter estimation and student model training. Similar to [30], starting with the assumption that all pseudo-groundtruth flows at position p, *i.e.,* $\{\mathbf{F}^i_{real}(l)(p)\}^m_{l=1}$, are equally likely, we learn parameters of the student model Θ_S. Then, we fix Θ_S and treat the student model's prediction as the correct flow $\mathbf{F}^i_{real}(p) = \hat{\mathbf{F}}^i_{real}(p)$ to estimate the error distribution for each position, and then use Eq. 3 to reweight the pseudo labels and so on.

Specifically, we design the following objective function for training $\mathcal{M}_{\Theta_S}$ and $\mathcal{R}_{\boldsymbol{\Sigma}}$:

$$\begin{aligned} &\mathcal{L}(\Theta_S, \boldsymbol{\Sigma}) \\ &= \frac{1}{h \times w} \sum_{p=1}^{h\times w} \{\mathcal{W}^i_u(l)(p) \cdot \mathcal{L}_{flow_u}(\hat{u}^i_{real}(l)(p), \check{u}^i_{real}(l)(p)) \\ &+ \mathcal{W}^i_v(l)(p) \cdot \mathcal{L}_{flow_v}(\hat{v}^i_{real}(l)(p), \check{v}^i_{real}(l)(p))\} \\ &+ \lambda \cdot \mathcal{L}_{error}(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}), \end{aligned} \tag{4}$$

where $\mathcal{L}_{flow_u}$ (or $\mathcal{L}_{flow_v}$) and $\mathcal{L}_{error}$ are losses for supervising the flow and error distributions respectively. $\mathcal{L}_{flow_u}$ (or $\mathcal{L}_{flow_v}$) is composed of a smooth loss $\mathcal{L}_{\mathbf{smooth}}$ [39] and a MSE loss $\mathcal{L}_{\mathbf{mse}}$.

For each image pair $\mathbf{P}^i_{real}$, we measure the agreement of the student's prediction $\hat{\mathbf{F}}^i_{real}(p)$ with the pseudo flow $\check{\mathbf{F}}^i_{real}(l)(p)$ (by teacher). To measure the difference between the error distribution and the true posterior error distribution, we use Kullback–Leibler divergence:

$$\begin{aligned} \mathcal{L}_{error}(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}) &= \sum_{i=1}^{N_{real}} \mathcal{D}_{\mathbf{KL}}(\boldsymbol{\Sigma} \parallel \hat{\boldsymbol{\Sigma}}) \\ &= \sum_{i=1}^{N_{real}} (\log \frac{\hat{\sigma}_i}{\sigma_i} - \frac{1}{2} + \frac{\sigma_i^2}{2\hat{\sigma_i}^2}), \end{aligned} \tag{5}$$

where $\hat{\boldsymbol{\Sigma}}$ means the true posterior distribution, *i.e.,* the empirical variance of the measurements correct $\mathbf{F}^i_{real}$ minus $\check{\mathbf{F}}^i_{real}(l)$. This step corresponds to the **M-step** in the EM algorithm [10].

Then, in the **E-step**, the true posterior distribution at iteration t, denoted as $\hat{\sigma}^t_i$, is estimated by assuming the student model's prediticons are equal to correct flows, and update the error distributions for the next iteration $(t+1)$ by using the following equation:

$$(\sigma^{t+1}_i)^2 = (\sigma^t_i)^2 + \varphi((\hat{\sigma}^t_i)^2 - (\sigma^t_i)^2), \tag{6}$$

Algorithm 1 : **Iterative optimization algorithm**

Input: $\{\mathbf{P}^i_{real}\}_{i=1}^{N_{real}}$, $\{\check{\mathbf{F}}^i_{real}(l)\}_{i=1}^{N_{real}}$

initialize Θ_S randomly
initialize Σ to be zero
initialize $\mathcal{W}_u$ and $\mathcal{W}_v$ to be 1
repeat
  for $i = 1$ to N_{real} **do**
    fix Σ_i, compute $\mathcal{W}^i_u(l)$ and $\mathcal{W}^i_v(l)$ by Eq. 3
    optimize Θ_S by Eq. 4
  end for
  for $i = 1$ to N_{real} **do**
    fix Θ_S, update Σ_i by Eq. 6
  end for
until convergence

Output: updated Θ_S

where φ is the step size. We repeat **E-step** and **M-step** until convergence. More details about the EM-based iterative training algorithm are detailed in Algorithm 1.

3.4. Confidence-aware Loss

To further alleviate the influence of errors on the pseudo flows during knowledge distillation, we introduce a *confidence-aware loss* (CAL) to replace the standard MSE loss in $\mathcal{L}_{flow}$ (Eq. 4). Our idea is to produce a "soft mask" C_i along with each pseudo flow $\check{\mathbf{F}}^i_{real}$ (by $\mathcal{M}_{\Theta_T}$) based on the consistency between pseudo flows $\check{\mathbf{F}}^{i,S}_{real}$ and $\check{\mathbf{F}}^{i,T}_{real}$, where $\check{\mathbf{F}}^{i,S}_{real}$ means the flow from source to target images while $\check{\mathbf{F}}^{i,T}_{real}$ is the flow from target to source images. Specifically, we compute the differences between $\check{\mathbf{F}}^{i,S}_{real}$ and $\boldsymbol{t}(\check{\mathbf{F}}^{i,T}_{real})$: $\Delta\mathbf{F}^i = \check{\mathbf{F}}^{i,S}_{real} - \boldsymbol{t}(\check{\mathbf{F}}^{i,T}_{real})$, where $\boldsymbol{t}(\cdot)$ is an operation that maps the estimated flow from target back to source. For a position p, its confidence score can be computed as: $C_i(p) = 1 - \frac{1}{1+e^{-b(\|\Delta\mathbf{F}^i(p)\|_2 - \tau)}}$, where $b = 50$ and $\tau = 0.08$ are controlling parameters. Thus, we replace the MSE loss $\mathcal{L}_{\mathbf{mse}}$ for flow in Eq. 4 by:

$$\begin{aligned}&\mathcal{L}_{\mathbf{cal}}(\hat{\mathbf{F}}^i_{real}, \check{\mathbf{F}}^i_{real}) \\ &= \frac{1}{h \times w} \sum_{p=1}^{h \times w} \{ \Big(\| \left(\hat{u}^i_{real}(p) - \check{u}^i_{real}(l)(p)\right) \|^2 \cdot C_i(p) \Big) \\ &+ \Big(\| \left(\hat{v}^i_{real}(p) - \check{v}^i_{real}(l)(p)\right) \|^2 \cdot C_i(p) \Big) \}.\end{aligned} \tag{7}$$

In the above equation (Eq. 7), the "soft mask" C_i is applied separately for each dimension of the flow, which ensures that the student model focuses on consistent matches during knowledge distillation.

4. Experiments

4.1. Experimental Settings

Datasets and Evaluation Metric: We perform extensive experiments on three most-widely used benchmarks: PF-PASCAL [56], PF-WILLOW [18] and SPair-71k [52]. Following the standard experimental protocol [39, 55, 56], we use the probability of correct keypoint (PCK) to measure the precision of overall assignment (α_{img} for PF-PASCAL, and α_{box} for PF-WILLOW and SPair-71k).

Training Objective: To show the generalizability of our PMD approach, we train our model using different degrees of supervision: **self/un-supervised**, **weakly-supervised** and **strongly-supervised** regimes:

- **Learning under self- (un-) supervision**. In this setup, we first train our teacher model using *synthetically* generated training pairs from PASCAL VOC [15] as in [55,60]. Then, for training the student model, we *randomly* generate pairs by using PF-PASCAL images [56] as the `train` set. In our experiments, we generate $60,000$ real pairs where there are only $2,911$ ($<5\%$) matched pairs.
- **Learning under weak supervision (only well-paired images)**. In the weakly-supervised learning setup, the teacher model is trained with synthetically generated training data (the same as the self-supervised setup). For training the student model, we use the well-paired PF-PASCAL [56] or SPair-71k images [52] as the `train` sets. Note that, in our setup, the only manual knowledge is that the given training pairs include object(s) from the same category. Supervision signals like category labels or object masks are not used in our approach.
- **Learning under strong supervision**. In this setup, keypoint matches are given for each training image pair, which are used to guide the generation of groundtruth semantic flow for each pair. The teacher model is first trained with such annotated `train` set. Then, the student model is trained with our PMD on the same `train` set. Note that, keypoint matches will be removed during the training process of the student model.

Detailed Training Settings: For training our probabilistic teacher model, we set $\alpha = 0.05$ and $\beta = 0.01$ in Eq. 2, and use the Adam optimizer to learn parameters. The initial learning rate is set to 3×10^{-5} and divided by 5 after 40 epochs (100 epochs in total). For jointly training $\mathcal{R}_{\mathbf{\Sigma}}$ and $\mathcal{M}_{\Theta_T}$, we first run the probabilistic teacher model on each real image pair multiple ($m = 4$) times to create a `train` set which includes multiple macthing hypotheses for each image. The weights of the student model's feature extractor are initialized from ResNet-101 [21] pretrained on ImageNet [11], while the remaining weights are randomly initialized. The parameters of probabilistic error modeling module are initialized to be zero. We set $\lambda = 0.01$ in Eq. 4, and all parameters are iteratively updated by using Algorithm 1. For learning Θ_S, we set the learning rate to 3×10^{-5}. For updating $\mathbf{\Sigma}$, we set φ to 0.01. A total of 40 rounds are performed to train the student model.

Table 1. Per-class PCK ($\alpha_{\text{box}} = 0.1$) on SPair-71k `test` [52]. The best scores in each group are highlighted in **bold**.

Sup.	Method	aero	bike	bird	boat	bot	bus	car	cat	cha	cow	dog	hor	mbik	pers	plnt	she	trai	tv	all
Weak.	WS-SA$_{\text{res101}}$ [56]	22.2	17.6	41.9	15.1	38.1	27.4	27.2	31.8	12.8	26.8	22.6	14.2	20.0	22.2	17.9	10.4	32.2	35.1	20.9
	NCN$_{\text{res101}}$ [57]	17.9	12.2	32.1	11.7	29.0	19.9	16.1	39.2	9.9	23.9	18.8	15.7	17.4	15.9	14.8	9.6	24.2	31.1	20.1
	SFNet$_{\text{res101}}$ [39]	26.9	17.2	45.5	14.7	38.0	22.2	16.4	55.3	13.5	33.4	27.5	17.7	20.8	21.1	16.6	15.6	32.2	35.9	26.3
	Ours$_{\text{res101}}$	26.2	18.5	48.6	15.3	38.0	21.7	17.3	51.6	13.7	34.3	25.4	18.0	20.0	24.9	15.7	16.3	31.4	38.1	**26.5**
Strong.	HPF$_{\text{res101}}$ [51]	25.2	18.9	52.1	15.7	38.0	22.8	19.1	52.9	17.9	33.0	32.8	20.6	24.4	27.9	21.1	15.9	31.5	35.6	28.2
	OT$_{\text{res101-FCN}}$ [48]	34.9	20.7	63.8	21.1	43.5	27.3	21.3	63.1	20.0	42.9	42.5	31.1	29.8	35.0	27.7	24.4	48.4	40.8	35.6
	Ours$_{\text{res101}}$	38.5	23.7	60.3	18.1	42.7	39.3	27.6	60.6	14.0	54.0	41.8	34.6	27.0	25.2	22.1	29.9	70.1	42.8	**37.4**

Table 2. Quantitative comparison with SOTAs on the PF-PASCAL `test` and PF-WILLOW in terms of average PCK.

	Sup. signal	Method	PF-PASCAL ($\alpha_{\text{img}} = 0.1$)	PF-WILLOW ($\alpha_{\text{box}} = 0.1$)
	–	PF$_{\text{HOG}}$ [18]	55.6	54.0
Un.	syn. pairs	A2Net$_{\text{res101}}$ [60]	70.8	68.8
		CNNGeo$_{\text{res101}}$ [55]	69.5	69.2
	syn. + random	**Ours**$_{\text{res101}}$	**80.5**	**73.4**
Weak.	pixel-wise masks	SFNet$_{\text{res101}}$ [39]	81.9	73.5
		GSF$_{\text{res101}}$ [28]	**84.5**	**75.8**
	image-level labels	FCSS$_{\text{VGG}}$ [33]	69.6	61.0
		WS-SA$_{\text{res101}}$ [56]	74.8	70.2
		NCN$_{\text{res101}}$ [57]	78.9	67.0
		RTNs$_{\text{res101}}$ [32]	75.9	71.9
		Ours$_{\text{res101}}$	**81.2**	**74.7**
Strong.	keypoint matches	UCN$_{\text{GoogLeNet}}$ [8]	55.6	54.0
		SCNet-AG+$_{\text{vgg16}}$ [19]	72.2	70.4
		Arbicon-Net$_{\text{res101}}$ [5]	77.3	-
		DFS$_{\text{res101}}$ [25]	86.0	75.0
		DCCNet$_{\text{res101}}$ [26]	82.3	73.8
		ANC-Net$_{\text{res101}}$ [42]	88.7	-
		HPF$_{\text{res101}}$ [51]	80.4	74.4
		Ours$_{\text{res101}}$	**90.7**	**75.6**

Testing Phase: During test time, the input image pair is simply forwarded through the trained student model to generate a full-resolution semantic correspondence (flow) map.

4.2. Main Results

PF-PASCAL [56] and PF-WILLOW [18]: Following the experimental protocol in [39,51,55,56], we train our student model on PF-PASCAL `train` set, and evaluate the performance on PF-PASCAL `test` set and PF-WILLOW. Table 2 reports the PCK values of 16 SOTAs and ours. For fair comparison, these SOTAs are divided into different groups by degree of supervision. Across multiple degrees of supervision, our approaches with different settings achieve better performance. On PF-PASCAL, our self-supervised model yields very competitive result (80.5%), which is significantly better than all approaches under the same level of supervision, and is even better than most of approaches with stronger supervisions. Our weakly supervised model achieves the PCK score of 81.2%, which is better than existing weakly supervised approaches except for those using additional (pixel-wise) mask annotations. Our strongly supervised model sets a new record of PCK score 90.7%. Figure 5 provides the visual samples of different models on PF-PASCAL. On PF-WILLOW, our approaches also set new records for different degrees of supervision. The comparisons clearly show that our PMD can 1) transfer and enhance the knowledge learned from synthetic data to a new model (*i.e.,* our un-/weakly-supervised models) and, 2) can be extended to incorporate stronger key-point supervision for better accuracy (*i.e.,* our strongly-supervised models).

Table 3. Ablation studies on PF-PASCAL `test`. "ST" means the static teacher; "PT" stands for the probabilistic teacher; "PSR" is the probabilistic supervision reweighting module; "CAL" denotes the confidence-aware loss.

	Method	Sup. signal	Train Data	PF-PASCAL ($\alpha_{\text{img}} = 0.1$)
Un.	Teacher model (Static)	synthetic pairs	PASCAL VOC	76.6
	Teacher model (Probabilistic)	synthetic pairs	PASCAL VOC	76.3
	Student model	ST	PF-PASCAL (**random**)	54.1
	Student model + PSR	PT (m=2)	PF-PASCAL (**random**)	70.3
	Student model + PSR	PT (m=4)	PF-PASCAL (**random**)	75.2
	Student model + PSR	PT (m=8)	PF-PASCAL (**random**)	75.5
	Student model + PSR + CAL	PT (m=4)	PF-PASCAL (**random**)	**80.5**
Weak.	Teacher model (Static)	synthetic pairs	PASCAL VOC	76.6
	Teacher model (Probabilistic)	synthetic pairs	PASCAL VOC	76.3
	Student model	ST	PF-PASCAL (**paired**)	74.3
	Student model + PSR	PT (m=2)	PF-PASCAL (**paired**)	78.2
	Student model + PSR	PT (m=4)	PF-PASCAL (**paired**)	80.8
	Student model + PSR	PT (m=8)	PF-PASCAL (**paired**)	80.8
	Student model + PSR + CAL	PT (m=4)	PF-PASCAL (**paired**)	**81.2**

SPair-71k [52]: We follow [38, 51] to train all models on SPair-71k `train` set, and perform evaluation on SPair-71k `test` set. As shown in Table 1, on the largest and most challenging benchmark SPair-71k, our weakly supervised model achieves PCK score of 26.5%, which is higher than all approaches under the same degree of supervision. Our strongly supervised model sets a new PCK record of 37.4%. The comparisons again demonstrate the superiority of our approach compared to existing approaches.

4.3. Ablation Experiments

Here, our ablation experiments are conducted by using our **un-supervised** and **weakly supervised** models.

Effectiveness of Probabilistic Model Distillation: To show the effectiveness of distilling knowledge from a probabilistic teacher model (PT), we provide a baseline student model which is distilled from a static teacher model (conventional model distillation) under different settings. As shown in Table 3, in all settings, we observe that the baseline student model is upper-bounded by its static teacher model. In contrast, our PMD approach enables the same student model to achieve better performance than its probabilistic teacher model (PT). In the **un-supervised** set-

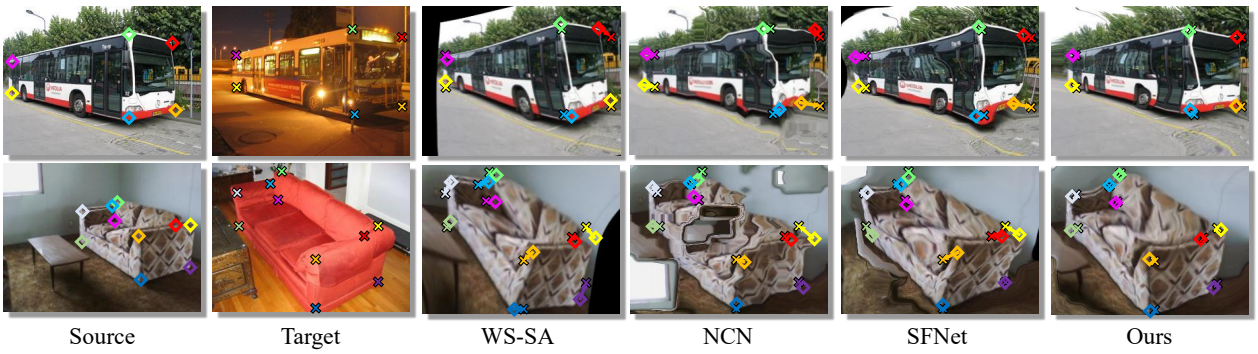

Figure 5. Quantitative comparison of dense correspondence. Diamonds and crosses denote the key points in source and target respectively, and vectors depict the matching error. *Zoom-in for details.*

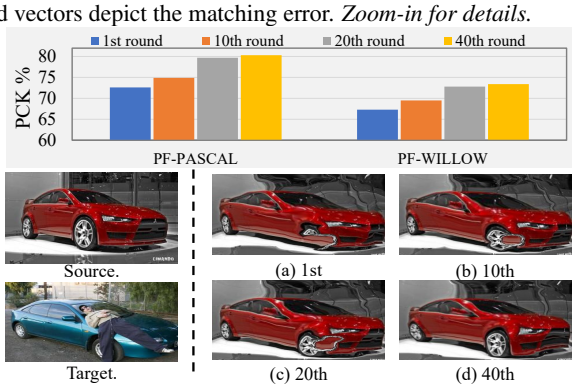

Figure 6. Performance analysis of each round. Top: Average PCK on two datasets; Bottom: Some visual examples.

ting, we observe that our confidence-aware loss (CAL) can greatly improve accuracy. It indicates that the "unmatched" pairs (*e.g.,* one image for a "dog" and the other for a "bus") can be filtered out by using our CAL loss. In the **weakly-supervised** setting, CAL still brings performance gain (0.4%). It can be observed that the probabilistic teacher model has a slight drop in performance compared to the static teacher model, yet it can provide meaningful pixel-wise hypotheses to better guide the learning of a student model. We also show that our PSR brings significant performance gains under both settings. Moreover, to verify the effectiveness of our iterative training algorithm, we provide comparisons of different rounds using our weakly-supervised model in Figure 6. Clearly, the average PCK scores gradually improve when more rounds of training are performed, which means the correct pixel-wise supervision signals can be gradually inferred during iterative training.

Effect of Unmatched Pairs: In the challenging unsupervised setting, the matched pairs only occupy less than 5% of the training data. Yet, our model achieves the PCK score of 80.5%, which is only 0.7% lower than our weakly supervised model (81.2%) on PF-PASCAL. A new question arises — *Does the proposed method really need any matched pairs*? To answer this question, we thoroughly investigate the effect of unmatched pairs. **First**, we fix the number of matched pairs to be $2,911$, and manually add different numbers of unmatched pairs to control the proportion of matched pairs. As can be seen in Figure 7 (a), our approach is almost immune to the image pair-level noise. This is because our PSR and CAL can best alleviate the influence of noise data. **Second**, we fix the number of unmatched pairs to be $60,000$, and manually add different numbers of matched pairs for training the student model. As shown in Figure 7 (b), the accuracy improves when more matched pairs are included for training. Therefore, we can conclude that our PMD approach requires matched pairs, yet it is robust to the image pair-level noise. It should be noted that, theoretically, matched pairs can be automatically generated if large amounts of (unlabeled) images are collected.

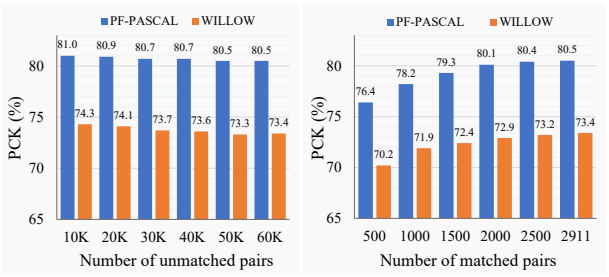

Figure 7. Analysis for the effect of unmatched pairs on the PF-PASCAL `test` set and PF-WILLOW in terms of average PCK.

5. Conclusion

In this paper, we introduce a novel probabilistic model distillation (PMD) approach, and use it within a probabilistic teacher-student framework to solve the semantic correspondence problem. The proposed method substantially outperforms existing self-supervised/un-supervised approaches, and even surpasses models trained with (strong) auxiliary annotations. Moreover, we show that our PMD can be extended to incorporate stronger supervision for better accuracy. The results demonstrate the effectiveness of our PMD approach for semantic correspondence. We believe that our PMD can be applied to other domains where annotations are difficult or labor-intensive to collect.

Acknowledgement: This research was funded by the NSFC (U1964203) and the National Key R&D Program Project of China (2017YFB0102603).

References

[1] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. In *NeurIPS*, 2014. 4

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *TOG*, 2009. 1

[3] John L Barron, David J Fleet, and Steven S Beauchemin. Performance of optical flow techniques. *IJCV*, 1994. 1

[4] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *NeurIPS*, 2017. 2

[5] Jianchun Chen, Lingjing Wang, Xiang Li, and Yi Fang. Arbicon-net: Arbitrary continuous geometric transformation networks for image registration. In *NeurIPS*, 2019. 2, 3, 7

[6] Yun-Chun Chen, Po-Hsiang Huang, Li-Yu Yu, Jia-Bin Huang, Ming-Hsuan Yang, and Yen-Yu Lin. Deep semantic matching with foreground detection and cycle-consistency. In *ACCV*, 2018. 2, 3

[7] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Show, match and segment: Joint weakly supervised learning of semantic matching and object co-segmentation. *TPAMI*, 2020. 2, 3

[8] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. In *NeurIPS*, 2016. 2, 3, 7

[9] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1

[10] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1977. 5

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3, 6

[12] Zhijie Deng, Yucen Luo, and Jun Zhu. Cluster alignment with a teacher for unsupervised domain adaptation. In *ICCV*, 2019. 2

[13] Thomas G Dietterich. Ensemble methods in machine learning. In *MCSW*, 2000. 4

[14] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, 2015. 3, 4

[15] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 6

[16] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *CVPR*, 2016. 2

[17] Yoav HaCohen, Eli Shechtman, Dan B Goldman, and Dani Lischinski. Non-rigid dense correspondence with applications for image enhancement. *TOG*, 2011. 1

[18] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow. In *CVPR*, 2016. 1, 3, 6, 7

[19] Kai Han, Rafael S Rezende, Bumsub Ham, Kwan-Yee K Wong, Minsu Cho, Cordelia Schmid, and Jean Ponce. Scnet: Learning semantic correspondence. In *ICCV*, 2017. 7

[20] Tal Hassner, Viki Mayzels, and Lihi Zelnik-Manor. On sifts and their scales. In *CVPR*, 2012. 2

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 6

[22] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv*, 2015. 1, 2

[23] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *ICANN*, 2011. 4

[24] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 3, 4

[25] Hao Huang, Jianchun Chen, Xiang Li, Lingjing Wang, and Yi Fang. Robust image matching by dynamic feature selection. *arXiv preprint arXiv:2008.05708*, 2020. 7

[26] Shuaiyi Huang, Qiuyue Wang, Songyang Zhang, Shipeng Yan, and Xuming He. Dynamic context correspondence network for semantic alignment. In *ICCV*, 2019. 2, 3, 7

[27] Phillip Isola and Ce Liu. Scene collaging: Analysis and synthesis of natural images with semantic layers. In *ICCV*, 2013. 1

[28] Sangryul Jeon, Dongbo Min, Seungryong Kim, Jihwan Choe, and Kwanghoon Sohn. Guided semantic flow. In *ECCV*, 2020. 7

[29] Sangryul Jeon, Dongbo Min, Seungryong Kim, and Kwanghoon Sohn. Joint learning of semantic alignment and object landmark detection. In *ICCV*, 2019. 2, 3

[30] Rong Jin and Zoubin Ghahramani. Learning with multiple labels. In *NeurIPS*, 2003. 5

[31] Jaechul Kim, Ce Liu, Fei Sha, and Kristen Grauman. Deformable spatial pyramid matching for fast dense correspondences. In *CVPR*, 2013. 1, 3

[32] Seungryong Kim, Stephen Lin, SANG RYUL JEON, Dongbo Min, and Kwanghoon Sohn. Recurrent transformer networks for semantic correspondence. In *NeurIPS*, 2018. 2, 3, 7

[33] Seungryong Kim, Dongbo Min, Bumsub Ham, Sangryul Jeon, Stephen Lin, and Kwanghoon Sohn. Fcss: Fully convolutional self-similarity for dense semantic correspondence. In *CVPR*, 2017. 7

[34] Seungryong Kim, Dongbo Min, Somi Jeong, Sunok Kim, Sangryul Jeon, and Kwanghoon Sohn. Semantic attribute matching networks. In *CVPR*, 2019. 1

[35] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv*, 2013. 4

[36] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017. 2

[37] Zakaria Laskar, Hamed Rezazadegan Tavakoli, and Juho Kannala. Semantic matching by weakly supervised 2d point set registration. In *WACV*, 2019. 2, 3

[38] Junghyup Lee, Dohyung Kim, Wonkyung Lee, Jean Ponce, and Bumsub Ham. Learning semantic correspondence exploiting an object-level prior. *arXiv*, 2019. 7

[39] Junghyup Lee, Dohyung Kim, Jean Ponce, and Bumsub Ham. Sfnet: Learning object-aware semantic correspondence. In *CVPR*, 2019. 2, 3, 4, 5, 6, 7

[40] Junsoo Lee, Eungyeup Kim, Yunsung Lee, Dongjun Kim, Jaehyuk Chang, and Jaegul Choo. Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence. In *CVPR*, 2020. 1

[41] Kimin Lee, Changho Hwang, Kyoung Soo Park, and Jinwoo Shin. Confident multiple choice learning. In *ICML*, 2017. 2

[42] Shuda Li, Kai Han, Theo W Costain, Henry Howard-Jenkins, and Victor Prisacariu. Correspondence networks with adaptive neighbourhood consensus. In *CVPR*, 2020. 2, 3, 7

[43] Xin Li, Fan Yang, Leiting Chen, and Hongbin Cai. Saliency transfer: An example-based method for salient object detection. In *IJCAI*, 2016. 1

[44] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. In *SIGGRAPH*, 2017. 1

[45] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *TPAMI*, 2010. 1, 2

[46] Ce Liu, Jenny Yuen, and Antonio Torralba. Nonparametric scene parsing via label transfer. *TPAMI*, 2011. 1, 3

[47] Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William T Freeman. Sift flow: Dense correspondence across different scenes. In *ECCV*, 2008. 1, 2

[48] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *CVPR*, 2020. 7

[49] Jonathan L Long, Ning Zhang, and Trevor Darrell. Do convnets learn correspondence? In *NeurIPS*, 2014. 3

[50] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 1

[51] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *ICCV*, 2019. 7

[52] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv*, 2019. 6, 7

[53] Weichao Qiu, Xinggang Wang, Xiang Bai, Alan Yuille, and Zhuowen Tu. Scale-space sift flow. In *WACV*, 2014. 1, 3

[54] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *CVPR*, 2018. 1, 2

[55] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *CVPR*, 2017. 2, 3, 4, 6, 7

[56] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. End-to-end weakly-supervised semantic alignment. In *CVPR*, 2018. 2, 3, 6, 7

[57] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *NeurIPS*, 2018. 2, 3, 7

[58] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015. 2

[59] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmentation in internet images. In *CVPR*, 2013. 1

[60] Paul Hongsuck Seo, Jongmin Lee, Deunsol Jung, Bohyung Han, and Minsu Cho. Attentive semantic alignment with offset-aware correlation kernels. In *ECCV*, 2018. 6, 7

[61] Berk Sevilmis and Benjamin B. Kimia. Alignment by composition. In *WACV*, 2019. 1, 3

[62] Jian Sun, Nan-Ning Zheng, and Heung-Yeung Shum. Stereo matching using belief propagation. *TPAMI*, 2003. 1

[63] Prune Truong, Martin Danelljan, and Radu Timofte. GLU-Net: Global-local universal network for dense flow and correspondences. In *CVPR*, 2020. 1, 3, 4

[64] Xiaolong Wang, Allan Jabri, and Alexei A. Efros. Learning correspondence from the cycle-consistency of time. In *CVPR*, 2019. 2, 3, 4

[65] Jeremy HM Wong and Mark Gales. Sequence student-teacher training of deep neural networks. In *ISCA*, 2016. 4

[66] Fan Yang, Xin Li, Hong Cheng, Jianping Li, and Leiting Chen. Object-aware dense semantic correspondence. In *CVPR*, 2017. 1, 3

[67] Hongsheng Yang, Wen-Yan Lin, and Jiangbo Lu. Daisy filter flow: A generalized discrete approach to dense correspondences. In *CVPR*, 2014. 1, 2

[68] Yang You, Chengkun Li, Yujing Lou, Zhoujun Cheng, Lizhuang Ma, Cewu Lu, and Weiming Wang. Semantic correspondence via 2d-3d-2d cycle. *arXiv*, 2020. 2, 3

[69] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017. 2

[70] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. Learning dense correspondence via 3d-guided cycle consistency. In *CVPR*, 2016. 2, 3