Learning Contrastive Representation for Semantic Correspondence

Taihong Xiao1 · Sifei Liu2 · Shalini De Mello2 · Zhiding Yu2 · Jan Kautz2 · Ming-Hsuan Yang1,3

Received: date / Accepted: date

Abstract Dense correspondence across semantically related images has been extensively studied, but still faces two challenges: 1) large variations in appearance, scale and pose exist even for objects from the same category, and 2) labeling pixel-level dense correspondences is labor intensive and infeasible to scale. Most existing methods focus on designing various matching modules using fully-supervised ImageNet pretrained networks. On the other hand, while a variety of self-supervised approaches are proposed to explicitly measure image-level similarities, correspondence matching the pixel level remains under-explored. In this work, we propose a multi-level contrastive learning approach for semantic matching, which does not rely on any ImageNet pretrained model. We show that image-level contrastive learning is a key component to encourage the convolutional features to find correspondence between similar objects, while the performance can be further enhanced by regularizing cross-instance cycle-consistency at intermediate feature levels. Experimental results on the PF-PASCAL, PF-WILLOW, and SPair-71k benchmark datasets demonstrate that our method performs favorably against the state-of-the-art approaches. The source code and trained models will be made available to the public.

Keywords Semantic Correspondence · Image-level Contrastive Learning · Pixel-level Contrastive Learning · Cross-instance Cycle Consistency

1 Introduction

Semantic correspondence is one of the fundamental problems in computer vision with many applications in object recognition (Duchenne et al. 2011; Liu et al. 2011), image editing (Dale et al. 2009), semantic segmentation (Kim et al. 2013), and scene parsing (Zhou et al. 2015a), to name a few. The goal is to establish dense correspondences across images containing the objects or scenes of the same category. For example, as shown in Fig. 1(a), such dense correspondence is established across two horse images, where the semantically similar keypoints such as eyes or ears of different horses are matched. However, this task is extremely challenging as different objects usually appear with distinctive appearances caused by variations in shapes, lighting, poses and scales. Classic approaches (Bristow et al. 2015; Hur et al. 2015; Kim et al. 2013; Liu et al. 2011; Taniai et al. 2016; Yang et al. 2014) determine correspondence matching via hand-crafted features such as SIFT (Lowe 2004), DAISY (Tola et al. 2010) and HOG (Dalal and Triggs 2005). More recently, deep CNN based approaches (Choy et al. 2016; Han et al. 2017; Jeon et al. 2018; Kanazawa et al. 2016; Kim et al. 2019; Novotny et al. 2017; Rocco et al. 2017, 2018a; Seo et al. 2018; Zhou et al. 2016) have achieved significant improvements, by exploiting hierarchical image features that provide rich semantic features that are invariant to intra-class variations.

It is challenging to develop fully-supervised approaches for learning pixel-level matching as a large number of im-
age pairs with detailed pixel correspondence annotations are required. To alleviate this issue, several methods exploit weakly-supervised information, e.g., object bounding boxes or foreground masks (Jeon et al. 2018; Kim et al. 2019; Zhou et al. 2015a, 2016; Lee et al. 2019), for this task. Nevertheless, it still entails significant amount of labor to annotate these labels for a large scale dataset. As a trade-off between model performance and labeling effort, several methods (Rocco et al. 2018a; Kim et al. 2018; Rocco et al. 2018b; Huang et al. 2019; Min et al. 2020) leverage image class labels as weak supervisory signals. For example, we can obtain image pairs from the same object or scene category, which provides weak supervision for learning the correspondence. However, existing methods predominantly rely on effective universal feature representations in the first place, which are often obtained by supervised ImageNet pretraining. Benefiting from large-scale labeled images, a supervised pretrained network can extract image-level discriminative features for semantic correspondence.

Recent advances (Hjelm et al. 2019; Chen et al. 2020; He et al. 2020; Grill et al. 2020) in self-supervised representation learning exploit contrastive loss functions to construct image-level discriminative features from unlabeled image data. Their goal is to push object representations of various views of the same object closer while pulling those of different objects further apart by learning to perform an image instance discrimination task. The learned representations by these contrastive learning methods have achieved significant performance gains for numerous pretext tasks such as video object segmentation (Pathak et al. 2017) and tracking (Wang and Gupta 2015). However, learning a fine-grained representation at part- or pixel-level has not been well studied, especially for semantic correspondence without using ImageNet pretrained models. Some methods (Wang et al. 2020; Pinheiro et al. 2020) have generalized image-level contrastive
learning to dense contrastive learning, but do not establish cross-instance correspondence. As shown in Fig. 1 and summarized in Table 1, existing approaches have used different levels of supervisory signals, while our model is learned based on weak supervision, i.e., image pairs of the same category, without resorting to fully-supervised pretrained ImageNet networks.

To learn a generalizable representation for semantic correspondence, we develop a multi-level contrastive representation learning method in this paper. We show that by applying the contrastive learning framework (He et al. 2020) merely on the image level, the mid-level convolutional features can capture local correspondences between similar objects reasonably well. The results suggest that to learn good representations for high-level tasks (e.g., object recognition), lower-level features are enforced to learn correct correlations at a fine-grained level as well. We embed a pixel-level contrastive learning scheme into the same network to further obtain fine-grained feature representation learning by enforcing cross-instance cycle consistency regularization at intermediate feature levels. Given a pair of images with semantically similar objects, we track selected pixels from the source to the target images and then back to the source via an affinity matrix. We then enforce that those selected pixels map back to their original locations via the self-attention module to localize the foreground object and cycle consistency is equivalent to pixel-level contrastive learning scheme. By relaxing the strong dependency on supervisory signals, our method is learned from early to late layers of a convolutional neural network. We pose a new weakly-supervised semantic matching problem by relaxing the strong dependency on supervised ImageNet pretrained models and removing the validation ground truth used for model selection.

We develop a multi-level contrastive learning framework where image-level contrastive learning generates object-level discriminative representations and pixel-level contrastive learning further facilitates fine-grained representations at region- or pixel-level to improve dense semantic correspondence performance.

We propose a cross-instance cycle consistency regularization to learn a discriminative local feature at the pixel-level without dense ground truth by leveraging different objects of the same category in different images instead of the same object in self-augmented images or video sequences.

We demonstrate that the proposed model performs favorably against the state-of-the-art method on three datasets with comprehensive ablation studies on components of our framework.

2 Related Work

2.1 Semantic Correspondence

Numerous methods exploit hierarchical features from deep models pretrained on the ImageNet dataset to infer semantic correspondence. In these approaches, semantic matching is formulated as a geometric alignment task and addressed via the self-supervised learning framework where training image pairs and ground truth are synthesized based on in-plane transformations (Jeon et al. 2018; Kanazawa et al. 2016; Kim et al. 2018; Rocco et al. 2017, 2018a; Han et al. 2017; Seo et al. 2018). On the other hand, weak-supervision signals, such as image-level labels (Rocco et al. 2018a; Kim et al. 2018; Rocco et al. 2018b; Huang et al. 2019; Min et al. 2020), bounding boxes (Jeon et al. 2018; Kim et al. 2019; Zhou et al. 2015a, 2016; Lee et al. 2019), and keypoints (Han et al. 2017) have also been used for semantic correspondence. In contrast, the proposed method does not require fully supervised ImageNet pretrained networks and only utilizes image-level supervision, i.e., image pairs of same category, to determine semantic correspondence from images.

In addition, numerous approaches freeze the ImageNet pretrained backbone model and determine the network structure of other modules or hyper-parameters by exploiting ground truth keypoint correspondences of a small validation set as supervisory signals (Min et al. 2019a; Liu et al. 2020; Huang et al. 2019; Chen et al. 2018). Min et al. (2019a) leverage a small number of relevant features selected from early to late layers of a convolutional neural network as well as beam search to construct hyperpixel layers, and use regularized Hough matching (RHM) to infer semantic correspondence efficiently. Liu et al. (2020) pose semantic matching as an optimal transport problem and solve it using the Sinkhorn’s algorithm (Sinkhorn 1967). These two methods require only supervision from keypoints of a small validation set during the beam search stage to select feature layers from a pretrained deep model (e.g., ResNet-50 and ResNet-101). We note that keypoint ground truth from a validation set provides strong supervision for model or hyper-parameter selection, which is not weaker than either pixel-level training supervision or supervised ImageNet pretrained model weights. In contrast, our method does not require any keypoint correspondence supervision from a validation dataset as the proposed cross-instance cycle consistency regularization can be used for self-supervised learning.
2.2 Self-Supervised Representation Learning

Self-supervised representation learning methods can be classified into three categories: generative, inductive (predictive), and contrastive. Generative models, e.g., (Vincent et al. 2008; Oord et al. 2016; Van Oord et al. 2016), maximize the likelihood of observed data based on various probabilistic formulations and representation, e.g., auto-encoders and convolutional neural networks. In contrast, inductive methods predict some known information from the data as there exist ground truths for comparison, which is typically carried out by applying various augmentation techniques, including image rotations prediction (Gidaris et al. 2018), spatial configuration of cropped patches (Doersch et al. 2015), video colorization (Zhang et al. 2016), video sequence order sorting (Misra et al. 2016), and jigsaw puzzle solving (Noroozi and Favaro 2016).

Recently, numerous contrastive learning methods, such as Deep InfoMax (Hjelm et al. 2019), SimCLR (Chen et al. 2020), MoCo (He et al. 2020) and BYOL (Grill et al. 2020), have shown that effective representation models can be constructed without supervision, with performance comparable to fully supervised ones. However, image-level contrastive learning may not produce optimal features for semantic correspondence, where the association is built between pixels. A few methods (Wang et al. 2020; Pinheiro et al. 2020; Jabri et al. 2020) propose to learn dense visual representations where the association is required between pixels, which are direct extensions of image-level instance contrastive learning methods (e.g., MoCo (He et al. 2020)). They compute contrastive losses between different views of the same instance, e.g., either via self-augmented images (Wang et al. 2020; Pinheiro et al. 2020), or from videos of the same instance (Jabri et al. 2020). However, their learned representations only captures the instance-level information, and thus cannot be used to establish the pixel-level cross-instance semantic correspondence. In contrast, we propose a multi-level cross-instance pixel contrastive learning method by leveraging different instances via a cyclic framework to infer semantic correspondence.

Kang et al. (2020) leverage the pixel-level cycle association of source and target pixel pairs across two different domains for contrastive representation learning. However, it requires knowing the pixel-level categories in advance to construct positive/negative samples for contrastive learning. In contrast, the proposed method does not entail any pixel-level supervisory signals to form positive/negative pixel pairs. On the other hand, Xie et al. (2021) propose a joint pixel-level and instance-level contrastive learning framework. It differs from our approach in constructing positive and negative samples. The pixel-level contrastive learning in (Xie et al. 2021) is carried out through pixels within the same instance, i.e., the corresponding pixels in two views of the same image are considered as positive pairs. Nevertheless, the proposed model establishes pixel-level contrastive learning across different instances. Extensive experimental results show that the proposed cross-instance pixel-level contrastively learned representation can better handle challenges caused by large appearance and pose distinctions across different instances.

2.3 Temporal Correspondence

Our work is related to temporal correspondence learning, for which a number of self-supervised approaches (Vondrick et al. 2018; Oord et al. 2018; Li et al. 2019; Wang et al. 2019; Jabri et al. 2020) have been explored. Li et al. (2019) propose to learn temporal correspondence by joining region-level localization and pixel-level matching through a shared inter-frame affinity matrix. In (Jabri et al. 2020), the correspondence learning problem is cast as link prediction in a space-time graph constructed from a video and long-range cycle consistency is exploited to learn temporal correspondence. Unlike these methods, we target a more challenging task for establishing correspondence across semantically similar instances with significant appearance and pose distinctions.

2.4 Cycle Consistency

Cycle consistency has been widely used as a constraint in numerous vision tasks. For example, in the context of image-to-image translation (Zhu et al. 2017), or face attributes editing (Zhou et al. 2017; Xiao et al. 2018a,b), exploiting cycle consistency enables learning the mapping between different image domains from unpaired data. In the context of optical flow estimation, computing the forward and backward consistency (Meister et al. 2018; Liu et al. 2019) can be used to effectively infer occluding pixels for learning optical flow.

For learning correspondence, the cycle consistency constraint is formulated in various ways. Zhou et al. (2016) construct a cross-instance loop between real and synthetic images and establish cross-instance correspondence through 3D CAD rendering. SC-Net (Han et al. 2017) establishes a differentiable flow field by computing feature similarities while considering background clutter and proposes flow consistency loss between forward and backward flow fields. Zhou et al. (2015b) propose to optimize joint matching of multiple images via rank minimization by translating cycle consistency into positive semi-definiteness and low-rankness constraints. On the other hand, Chen et al. (2018) exploit forward-backward consistency and transitivity consistency constraints to enforce geometrically plausible predictions. In contrast, our cycle consistency regularization...
is established via self augmentations to learn finer-grained pixel-level representations.

3 Proposed Method

In this section, we introduce the proposed contrastive representation learning model for semantic correspondence. As shown in Fig. 2, the overall framework consists of image-level and pixel-level contrastive learning modules. We briefly describe the image-level contrastive learning schemes in Section 3.1, introduce the proposed pixel-level contrastive learning method in Section 3.2 and our implementation details in Section 3.3.

3.1 Image-level Contrastive Learning

Image-level contrastive learning (He et al. 2020; Chen et al. 2020; Grill et al. 2020) aims to extract object-level discriminative features in a self-supervised manner. It learns representations from object-centric images by minimizing the distance between two different views of an object generated through different augmentation methods. In addition, numerous methods (He et al. 2020; Chen et al. 2020) use negative samples from different images and maximize the distance between positive and negative observations. In this work, we exploit negative examples in a way similar to the MoCo method (He et al. 2020). Image-level contrastive learning can be considered as a dictionary look-up task. We use a dynamic queue of size $K$ to store a set of encoded keys \{f_1, f_2, \ldots\}, among which a single positive key $f_k$ matches with $f_q$. For a query $f_q$, its positive key $f_k$ encodes a different view of the same image, while the negative keys encode the views of different images. Thus, the image-level contrastive loss is defined as follow:

$$L_q = -\log \frac{\exp(f_q \cdot f_k/\tau)}{\sum_{j \neq k} \exp(f_q \cdot f_j/\tau)}, \quad (1)$$

where $\tau$ denotes a temperature hyper-parameter. Both the query network $q$ and the key network $k$ share the same architecture, but $q$ is updated based on the image-level contrastive loss using back-propagation while $k$ is updated with a momentum function:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q, \quad (2)$$

where $\theta_q$ and $\theta_k$ are parameters of $q$ and $k$, and $m \in [0, 1)$ denotes the momentum coefficient.

For image-level contrastive learning, each image is regarded as a unique category. Therefore the function for image-level contrastive learning approaches (1) is essentially equivalent to a $(K + 1)$-way classification task, where $K$ is the size of the queue. It facilitates learning image-level discriminative representations without using any image labels. We show in our experiments (see Table 3) that even without additional objective functions, the learned image-level representation models can perform semantic correspondence matching well (see Fig. 3).

3.2 Pixel-level Contrastive Learning

To learn fine-grained representation models, we propose a pixel-level contrastive learning method explicitly for semantic correspondence. As shown in Fig. 2(b), a pair of images $I_0$ and $I_1$ containing different objects of the same category are provided as the inputs. We then obtain an attention map highlighting the foreground object region via the self-attention module. Based on the attention map, we obtain $I_0'$ by applying a series of spatial data augmentations to $I_0$, including random horizontal flipping, rotation, and cropping. We use the same encoder network $q$, shared with the image-level module, to extract features for $I_0$, $I_1$, and $I_0'$, which are denoted by $F_0$, $F_1$ and $F_0'$, respectively, as shown in Fig. 2(b). We extract feature maps from the middle layers of ResNet50 instead of the feature vector from the last layer to learn the pixel-level representation. We introduce each sub-module in the following sections.

Self-attention module. We introduce a self-attention module to avoid including pixels from background regions. For the given source image $I_0$, we extract the feature $f_0 \in \mathbb{R}^C$ after the last layer and the feature $K_0 \in \mathbb{R}^{C \times H \times W}$ before the global average pooling layer, where $H$ and $W$ denote the size of the feature map. As the feature map $K_0$ keeps the spatial location information of pixels in $I_0$, we compute cosine similarity between normalized $f_0$ and all feature vectors in the normalized attention map $K_0$. The self-attention module is used to exclude background pixels when applying a series of augmentation to $I_0$ to obtain $I_0'$. The random image regions are cropped near the pixels using the largest cosine similarity. As such, the background pixels are less likely to be contained in the augmented image $I_0'$, thus eliminating the number of pixel paths starting from the background for more effective cross-instance cycle consistency.

Associating pixels via affinity. Given a pair of images $I_0$ and $I_1$ and their mid-level features $F_0 \in \mathbb{R}^{C \times H_0 \times W_0}$ and $F_1 \in \mathbb{R}^{C \times H_1 \times W_1}$, we use the correlation matrix $R_{01} \in \mathbb{R}^{R_0 \times R_1}$ to represent the pixel-level similarity as:

$$R_{01} = F_0^\top F_1, \quad (3)$$

To ensure a one-to-one mapping, the correlation matrix needs to be sparse. However, it is challenging to model a sparse matrix in a deep neural network. Therefore, we relax this constraint and encourage the correlation matrix to be
sparse by normalizing each row with the softmax function. As such, the similarity score distribution can be peaked, and only a few pixels with high similarity in the source image are matched to each point in the target image. The affinity matrix is defined by:

\[ A_{01} = \text{softmax}(R_{01}/t), \]  

where \( t \) is the temperature hyperparameter controlling how peaked the normalized distribution is. The affinity matrix enjoys several good properties: 1) The summation over each row is unity since softmax is applied to the row dimension. 2) The multiplication of two affinity matrices results in an affinity matrix. 3) The affinity matrix can be used to trace the corresponding pixel locations of feature map \( F_0 \) in the target feature map \( F_1 \), defined by \( \hat{P}_{01} = A_{01}G_1 \), where \( G_1 \in \mathbb{R}^{H_1 \times W_1 \times 2} \) is a vectorized pixel location map (i.e., each element denotes its horizontal and vertical positions).

**Pixel-level guidance.** We carry out the proposed pixel-level contrastive learning by formulating the correspondence as a graph, where the nodes are image pixels and the edges are weighted by the similarities between their features in the latent space. Starting from \( I_0' \) to \( I_1 \) and then back to \( I_0 \), we track corresponding pixels by computing two affinity matrices \( A_{01}' \) and \( A_{10} \) following (4) using their mid-layer features. We enforce cross-instance cycle consistency by requiring pixels from \( I_0' \) to be mapped back to where they are located in \( I_0 \), through a different image \( I_1 \). The affinity matrix along the path of cycle can be simply obtained through multiplication, \( A_{01} = A_{01}'A_{10} \in \mathbb{R}^{H_0W_0 \times H_1W_1} \), where \( H \) and \( W \) denote the feature size. Each element in the cycle affinity matrix \( A_{01} \) depicts a pixel path from \( I_0' \) to \( I_0 \) passing through \( I_1 \). We predict corresponding pixel locations of the patch image \( I_0' \) in the source image \( I_0 \) by \( \hat{P} = A_{01}'G_0 \). As the ground truth correspondence \( \hat{P} \) between \( I_0' \) and \( I_0 \) is known, the pixel-level contrastive loss is defined as:

\[ L_p = ||P - \hat{P}||_2. \]  

We note that (5) is a variant of the contrastive objective w.r.t. pixels from a pair of images. Unlike the image-level loss (1) that explicitly pushes and pulls on the positive and negative pairs, the cycle loss matches a group of “starting” and “ending” pixels. Specifically, all pixels sampled in the walking path are considered positive, while the other pairs are negative. The feature representation is learned to be pixel-wisely discriminative when the affinity matrix is enforced to be “peaky” in each row.

**Information entropy regularization.** Information entropy loss (Min et al. 2020) can also be used as a regularization term to learn affinity by encouraging more distinctive correspondences. We empirically find that training the network by the pixel-level contrastive loss with the information entropy loss can further improve the performance of semantic correspondence. With the correlation matrix \( R \in \mathbb{R}^{H_0W_0 \times H_1W_1} \) as defined in (3), we compute the correlation entropy as:

\[ H(R) = -\frac{1}{H_0W_0} \sum_{i=1}^{H_0W_0} \sum_{j=1}^{H_1W_1} \phi(R)_{ij} \log \phi(R)_{ij}, \]  

where \( \phi(\cdot) \) denotes row-wise \( \ell_1 \) normalization, and \( \phi(R)_{ij} \) denotes the \( (i, j) \)-th elements of \( \phi(R) \). As lower correlation entropy indicates more distinctive correspondence between two images, we encourage low entropy for the image pair \( I_0 \) and \( I_1 \) by using the following information entropy loss:

\[ L_r = H(R_{01}) + H(R_{10}), \]  

where \( R_{01} \) and \( R_{10} \) are correlation matrices between the source and target images. We use correlation matrices \( R \)
instead of affinity matrices $A$ to compute the information entropy loss. The reason is that each row in $A = \text{softmax}(R/t)$ has been normalized with summation of each row to 1, and thus all entries would have been normalized twice if we use $\phi(A)$.

### 3.3 Implementation Details

**Pretraining and training.** We use ResNet50 (He et al. 2016) as the backbone network for feature extraction. For image-level contrastive learning, we use the feature from the last layer of ResNet50, while the feature from an intermediate layer (e.g., the feature after 13th Res-block) for pixel-level contrastive learning. Following the same hyperparameter settings of MoCo (He et al. 2020), we pretrain the backbone network using image-level contrastive learning on the ImageNet dataset for initialization. Then we train the backbone network with a combination of image-level and pixel-level contrastive losses along with the information entropy regularization,

$$L = \lambda_p L_p + \lambda_q L_q + \lambda_r L_r,$$

where $\lambda_p$, $\lambda_q$, and $\lambda_r$ are weight coefficients for the pixel-level contrastive loss (5), image-level contrastive loss (1) and the information entropy regularization (7), respectively. Note that for image-level contrastive learning, we do not need a 1000-class labeled dataset, but just random images. Neither do we need that for pixel-level contrastive learning. In contrast, only image pairs from the same category are required for pixel-level contrastive learning. We empirically set $\lambda_p = 0.0005$, $\lambda_q = 1$, $\lambda_r = 0.001$, and the temperature $t = 0.0007$ in (4).

**Validation: beam search without ground truth correspondence.** To utilize the learned representation model for semantic correspondence, we need to choose multi-layer features for testing, which is also called hyperpixel construction (Min et al. 2019a). Existing methods (Min et al. 2019a; Liu et al. 2020) use the beam search algorithm to find the optimal subset of deep convolutional layers according to performance on the validation split. However, the existing beam search method requires accessing the ground truth correspondence of the validation set. To relax the dependency on validation annotations, we perform beam search over all convolutional layers of a given deep model by using the proposed pixel-level contrastive loss as the performance indicator. A lower pixel-level contrastive loss means a better layer combination. The validation process is only used for hyperpixel selection and not for other hyper-parameter selection.

### 4 Experiments and Analysis

We propose a new benchmark setting specifically for learning semantic correspondence without using any supervised ImageNet pretrained network or validation ground truth. For all the evaluated methods, we use the ResNet50 model (He et al. 2016) as the backbone network for feature extraction, in a way similar to the self-supervised pretraining scheme of (He et al. 2020). We also use a unified standard for the validation step by performing model selection without resorting to using the ground truth keypoint correspondence of the validation image pairs.

In Section 4.1, we first describe the datasets and evaluation metrics. Then, we present results as per the new benchmark setting by comparing our method with state-of-the-art methods in Section 4.2. In Section 4.3, we conduct comprehensive ablation studies to analyze different components and variations of our model.

#### 4.1 Datasets and Evaluation Metrics

We evaluate the proposed method on three benchmark datasets: PF-PASCAL (Ham et al. 2018), PF-WILLOW (Ham et al. 2016) and SPair-71k (Min et al. 2019b). The PF-PASCAL dataset contains 1,351 image pairs from 20 object categories of the PASCAL VOC (Everingham et al. 2014) database, and the PF-WILLOW dataset contains 900 image pairs of 4 object categories. The SPair-71k dataset is a more challenging large-scale dataset consisting of keypoint-annotated 70,958 image pairs from 18 categories with diverse view-point and scale variations. We carry out pixel-level contrastive learning on one of these
Fig. 3: Visualization of the baseline and our method on the PF-PASCAL dataset. The first three columns show the correspondence predictions made based on the affinity matrix (Raw), transport matrix (after OT) and voting matrix (after OT + RHM). The last column is the ground truth correspondence. Different colors indicate different keypoint matches.
Table 2: Evaluation results on the PF-PASCAL and PF-WILLOW datasets. We compare our method against others with three different thresholds. Based on whether we add the entropy loss Eq. (7) or not, there are two variants of method, denoted by w/o and w/ entropy in the table. Numbers in bold indicate the best performance.

| Methods                  | PF-PASCAL ($\alpha_{img}$) | PF-PASCAL ($\alpha_{bbox}$) | PF-WILLOW ($\alpha_{bbox}$) |
|--------------------------|-----------------------------|------------------------------|-----------------------------|
|                          | PCK@0.05 | PCK@0.10 | PCK@0.15 | PCK@0.05 | PCK@0.10 | PCK@0.15 |
| Weakalign (Rocco et al. 2018a) | 30.83    | 60.90    | 76.49    | 27.98    | 56.29    | 72.42    |
| NC-Net (Rocco et al. 2018b)      | 41.34    | 62.89    | 72.14    | -        | -        | -        |
| DCC-Net (Huang et al. 2019)       | 45.28    | 72.26    | 82.43    | 37.29    | 65.01    | 78.10    |
| DHPF (Min et al. 2020)            | 45.64    | 73.96    | 85.69    | 43.09    | 69.01    | 82.14    |
| SCOT (Liu et al. 2020)            | 44.30    | 71.20    | 83.80    | 37.20    | 62.40    | 75.90    |
| Ours (w/o entropy)               | 51.00    | 77.10    | 88.50    | 40.10    | 66.40    | 80.30    |
| Ours (w/ entropy)                | 50.70    | 76.10    | 85.80    | 43.10    | 69.80    | 82.40    |

Three datasets and image-level contrastive learning method on the ImageNet dataset (Deng et al. 2009).

The PF-PASCAL dataset consists of 1300 image pairs with keypoint annotations of 20 object classes. Each pair of images in the PF-PASCAL dataset share the same set of non-occluded keypoints. We divide the dataset into 700 training pairs, 300 validation pairs, and 300 testing pairs following (Min et al. 2020; Liu et al. 2020). The image pairs for training/validation/testing are distributed proportionally to the number of image pairs of each object class.

For quantitative evaluation, we adopt the widely-used percentage of correct keypoints (PCK) metric, which counts the number of correctly predicted keypoints given a fixed threshold. Given a predicted keypoint \( k_{pr} \) and the ground truth keypoint \( k_{gt} \), the prediction is considered correct if:

\[
d(k_{pr}, k_{gt}) \leq \alpha \cdot \max(w_{\tau}, h_{\tau}),
\]

where \( d(\cdot, \cdot) \) is the Euclidean distance, \( w_{\tau} \) and \( h_{\tau} \) are the width and height of either an entire image or object bounding box, i.e., \( \tau \in \{img, bbox\} \), and \( \alpha \) is a fixed threshold. We compute the final PCK by averaging the results from all testing image pairs.

4.2 Evaluation against the State-of-the-art Methods

Due to the full supervision and richer features of deeper neural networks (e.g., ResNet101), models that are pretrained with the large-scale labeled ImageNet have stronger representation strength. For fair comparisons, we re-implement all the evaluated methods by initializing their models with the same MoCo (He et al. 2020) pretrained ResNet50 backbone network. We emphasize that the backbone networks weights of NC-Net (Rocco et al. 2018b), DCC-Net (Huang et al. 2019) and DHPF (Min et al. 2020) are all frozen during training as stated in the experimental sections in these papers. The SCOT method (Liu et al. 2020) does not involve any training. Instead, it adopts a fixed backbone and optimizes on top of the extracted features. Since all the methods are built on top of a fixed pretrained network, it is fair to evaluate all methods by replacing their backbones with the same MoCo pretrained ResNet50 model. Replacing these models with purely randomly initialized weights will degrade the performance. Note that we specifically evaluate the methods that utilized the same level of weak supervision as ours, i.e., a group of matchable image pairs.

As aforementioned, the ground truth annotations of the validation set are utilized by numerous approaches, e.g., Weakalign (Rocco et al. 2018a), NC-Net (Rocco et al. 2018b), DCC-Net (Huang et al. 2019), DHPF (Min et al. 2020), to determine the optimal hyper-parameters (neighborhood numbers, kernel size, etc.). For these method, we adopt the hyper-parameters in their original implementation in our re-implementations, because other randomly selected hyper-parameters would lead to lower performance. In addition, we utilize the re-implemented SCOT (Liu et al. 2020) method as our baseline model, which shares a similar matching procedure, i.e., OT and RHM, but with features extracted from a fixed backbone network pretrained with the image-level contrastive loss. For our implementation of SCOT, we remove the use of the validation set for beam search and replace it with the same strategy as proposed in Section 3.3.

Table 2 shows improvements introduced by our proposed multi-level contrastive learning method. We also visualize the results of the re-implemented SCOT baseline model and our method in Fig. 3. Our method generates more accurate and consistent matches than the baseline model (see more discussion of the reasons accounting for the improvement in the ablation study below). The re-implemented SCOT baseline model shows that the intermediate features of the backbone network pretrained using image-level contrastive learning can capture the semantic correspondences well. Furthermore, by embedding pixel-level contrastive learning into the network, the proposed model can adapt to the variations between different images at the pixel level, thereby generating more fine-grained matching results.
4.3 Ablation Studies

In this section, we analyze numerous components of the proposed method, including the image-level and pixel-level contrastive losses, beam search, OT and RHM, information entropy regularization, self-attention module, and the training layers. As each component plays an essential role in the model’s performance, we conduct comprehensive ablation studies and discuss the contribution of each module.

**Variants of image- and pixel-level contrastive learning.** We analyze the proposed pixel-level contrastive learning with comparisons to other possible variants on the PF-PASCAL dataset, as summarized below and in Table 3. For simplicity, we use IC and PC as the abbreviations of image-level and pixel-level contrastive learning variants:

1. IC: We re-implement SCOT as our baseline model, where the backbone is pretrained via image-level contrastive learning, as aforementioned in Section 4.2.
2. IC (finetune): To verify whether the performance gain above the baseline model comes from the image pairs in the PF-PASCAL dataset or not, we finetune the baseline model via the image-level contrastive loss on both ImageNet and PF-PASCAL datasets.
3. IC + PC (selfcycle): In addition to image-level contrastive learning, we conduct pixel-level contrastive learning on self-augmented image pairs from ImageNet.
4. IC + PC (align): We apply image- and pixel-level contrastive learning on both ImageNet and PF-PASCAL datasets, where the pixel-level contrastive loss of ImageNet is calculated on the self-augmented image pairs.
5. IC (fix) + PC: To verify the effectiveness of multi-level contrastive learning, we fix the backbone pretrained model via image-level contrastive learning while training a side convolutional net branch combining features from multiple layers as input via the pixel-level contrastive loss. The integrated feature following these additional layers is directly used for inference without applying beam search to the trained model.
6. IC + PC (ours): We jointly conduct image-level contrastive learning (on ImageNet) and pixel-level contrastive learning (on PF-PASCAL).

As shown in Table 3, the proposed IC + PC model outperforms the IC baseline method by a large margin (51.0% vs. 44.3%). Either the effectiveness of pixel-level contrastive learning or using image pairs from the same category accounts for the performance gain. To better understand these effects, we conduct the IC (finetune) experiment. As it performs almost as well as the IC variant (43.7% vs. 44.3%), the performance gain of the IC + PC model does not result from the weak supervision provided by the PF-PASCAL dataset. In other words, the proposed pixel-level contrastive learning method helps achieve the performance gain.

Results of IC + PC (selfcycle) and IC + PC (align) show that using augmented image pairs from ImageNet for pixel-level contrastive learning negatively affects performance (38.1% and 34.6% vs. 44.3%). The underlying reason is that different views of the same object may lead to trivial solutions as such augmented image pairs cannot provide views of different objects with a large variation. The result of IC (fix) + PC model is worse than that of the IC variant (39.5% vs. 44.3%), which shows that model performance degrades without leveraging the image-level contrastive loss for updating the model weights. It also reflects that the image-level contrastive learning scheme is the cornerstone of the pixel-level contrastive learning scheme. The representations learned via the image-level contrastive learning scheme lead to coarse correspondences, and they can be refined via the pixel-level contrastive learning scheme, leading to more accurate correspondences.

**Beam search, OT and RHM.** To validate the effectiveness of adopting the pixel-level contrastive loss as the objective in beam search (Section 3.3), we compare the results by performing beam search with and without using the ground truth annotations in the validation set. As shown in Table 4, beam search using only the pixel-level contrastive loss performs effectively in our method, as it gives nearly the same hyperpixel selection as the standard beam search algorithm (51.0% vs. 53.4%). This also reveals that the proposed pixel-level contrastive loss, as an unsupervised surrogate loss, facilitates learning effective fine-grained feature representations for cross-instance matching. In addition, we also study the effectiveness of OT and RHM. The results in Table 4 show that both OT and RHM effectively facilitate the matching process. More interestingly, even without using these two post-processing steps, our method achieves significant performance gain compared to the baseline ap-

---

**Table 3: Ablation study of the image- and pixel-level contrastive losses.** All the numbers are evaluated at PCK@0.05 on the PF-PASCAL dataset. IC and PC are abbreviations of image-level contrastive learning and pixel-level contrastive learning. IN and PF stand for the ImageNet and PF-PASCAL datasets. Please refer to details of the ablation studies described in Section 4.3 for details.

| Models               | IC | PC | PCK @0.05 |
|----------------------|----|----|-----------|
| IC                   | ✓  |    | 43.5      |
| IC (finetune)        | ✓  | ✓  | 43.7      |
| IC + PC (selfcycle)  | ✓  | ✓  | 38.1      |
| IC + PC (align)      | ✓  | ✓  | ✓  | 34.6      |
| IC (fix) + PC        | ✓  | ✓  | ✓  | 39.5      |
| IC + PC (Ours)       | ✓  | ✓  | 51.0      |
Table 4: Ablation study of beam search, optimal transport (OT), and regularized Hough matching (RHM). The w/o GT in the second column denotes the proposed beam search by using the pixel-level contrastive loss as the indicator. The numbers in the third column indicate the res-block IDs used for hyperpixel construction. The columns of OT and RHM denote whether they were used in the testing phase. The numbers in the last column are evaluated at PCK@0.05 on the PF-PASCAL dataset.

| Beam Search | Hyperpixel | OT | RHM | PCK @0.05 |
|-------------|------------|----|-----|-----------|
| w/o GT      | (1,2,4,13,14) | ✓  | ✓   | 44.3      |
| w/o GT      | (1,2,4,13,14) | ✓  | ✓   | 20.9      |
| w/o GT      | (1,2,4,13,14) | ✓  | ✓   | 18.8      |
| w/ GT       | (3,11,12,13,15) | ✓  | ✓   | 52.4      |
| w/ GT       | (3,11,12,13,15) | ✓  | ✓   | 39.1      |
| w/ GT       | (3,11,12,13,15) | ✓  | ✓   | 34.3      |
| w/o GT      | (2,12,13,15) | ✓  | ✓   | 51.0      |
| w/o GT      | (2,12,13,15) | ✓  | ✓   | 45.0      |
| w/o GT      | (2,12,13,15) | ✓  | ✓   | 41.0      |
| w/ GT       | (2,11,12,15) | ✓  | ✓   | 53.4      |
| w/ GT       | (2,11,12,15) | ✓  | ✓   | 46.7      |
| w/ GT       | (2,11,12,15) | ✓  | ✓   | 41.3      |

Table 5: Ablation study of the self-attention module on PF-PASCAL. We evaluate the performance at PCK@0.05.

| Ours | Self-Attention | PCK@0.05 |
|------|---------------|----------|
| Ours (w/o attention) | ✓ | 51.0 |
| Ours | ✓ | 49.9 |

Table 6: Evaluation results of using features from different layers on PF-PASCAL. The first column denotes the residual block ID where the feature for pixel-level contrastive learning is extracted. The second column represents the hyperpixel used for testing our beam search algorithm.

| Block ID | Hyperpixel | PCK@0.05 |
|----------|------------|----------|
| 13       | (2,12,13,15) | 51.0     |
| 12       | (2,6,13,15)   | 49.1     |
| 10       | (2,6,13,15)   | 52.0     |
| 9        | (2,6,13,15)   | 51.3     |
| 7        | (2,6,13,15)   | 49.1     |

Those background pixels in the target image should not be matched to those in the wall regions of $I_0$ since they do not belong to the same category. The self-attention module can help remove the false positive samples in the background to achieve better results. More visualization of the attention map can be found in Fig. 4.

Different layers of feature for training. We explore the effectiveness of features from different layers for pixel-level contrastive learning. For example, the ResNet50 model contains 16 res-blocks. We classify features before the 7th block as low-level features, those after the 13th block as high-level features, and the rest as middle-level features. Since features from shallow layers usually entail low-level vision clues, e.g., color and texture, and high-level layer features are invariant to different local regions (i.e., via image-contrastive loss), we only evaluate the mid-level features. Table 6 shows that the features from layer 7 to 13 help achieve similar performance while the 10th layer yields the best features. Even so, we utilize the features following the 13th layer for pixel-level contrastive learning in the other experiments to fairly compare with other state-of-the-art methods.

Differentiable OT, RHM, and concentration loss (Li et al. 2019). We observe that both OT and RHM effectively improve the model performance when used as post-processing steps in the testing pipeline. Motivated by this, we analyze whether they can improve the performance in our model when including them in the training phase. We implement both OT and RHM as differentiable layers without trainable parameters. The function of the differentiable OT layer is equivalent to the Hungarian algorithm (Munkres 1957), classically used for bipartite matching. We apply one iteration of the Sinkhorn’s algorithm to the affinity matrix...
Table 7: Ablation study of differentiable OT, RHM, and concentration loss on PF-PASCAL. OT and RHM are used in training as differentiable layers. The concentration loss is employed as a regularization during training. The results below are evaluated at PCK@0.05, and the bold number indicates the best performance.

| OT | RHM | Concentration | PCK@0.05 |
|----|-----|---------------|----------|
| ✓  | ✓   |               | 51.0     |
| ✓  |                      | 32.4     |
| ✓  |                      | 41.5     |
| ✓  |                      | 48.1     |

and obtain the transport matrix. Differentiable RHM aims to achieve geometrically consistent matching by re-weighting the matching scores and outputs the voting matrix. Here we insert the differentiable OT and RHM layers into the training pipeline and compute the pixel-level contrastive loss on either transport matrix (after OT) or voting matrix (after OT and RHM) instead of the affinity matrix. In addition, we adopt the concentration loss (Li et al. 2019) to encourage more concentrated keypoints predictions for the patch image.

By adopting differentiable OT and/or RHM layers, we observe an unstable training process. The results also reveal that the differentiable OT and RHM layers negatively affect the representation learning performance (Table 7). This is potentially caused by the distorted feature space. The gradients become unstable because the differentiable OT is carried out by iterative matrix inversion. We also observe that the matched pixels are already concentrated (see Fig. 3), and thus adding the concentration regularization does not make significant differences.

4.4 Discussion and Limitation

To better understand how the pixel-level contrastive loss function regularizes the feature representation, we visualize the correspondences given by the affinity matrices on the training dataset. For example, let $I_0$ and $I_1$ denote source and target images, and $I'_0$ the patch image randomly cropped from $I_0$. Fig. 4 shows that the cycle constraint can help the cross-instance correspondence prediction of $I'_0$ onto $I_1$ by enforcing the correspondence predictions of $I'_0$ on $I_0$ to be the same as the ground truth. For example, the patch image $I'_0$ in the fourth row is cropped from the right part of a juice bottle. Interestingly, the correspondence prediction on $I_1$ (the second column) is also on the right side of a different bottle. We observe a similar phenomenon in the other rows. These results demonstrate that multi-level contrastive learning helps focus on the contextual semantic information and learn spatially aware pixel-level feature representations.

Table 8: Evaluation results on the SPair-71k dataset. Numbers in bold indicate the best performance.

| Methods       | PCK@0.05 | PCK@0.10 | PCK@0.15 |
|---------------|----------|----------|----------|
| SCOT (Liu et al. 2020) | 6.20 | 15.60 | 24.60 |
| Ours          | 6.30     | 15.80    | 25.20    |

Table 9: Evaluation results on the SPair-71k dataset by using different percentages of pixel correspondence ground truth. For example, 5% GT means that 5% of training image pairs are provided with their pixel correspondence ground truth.

| Methods       | PCK@0.05 | PCK@0.10 | PCK@0.15 |
|---------------|----------|----------|----------|
| Ours (5% GT)  | 6.58     | 16.03    | 25.41    |
| Ours (10% GT) | 8.51     | 19.91    | 30.20    |
| Ours (15% GT) | 9.75     | 22.49    | 33.56    |
| Ours (100% GT)| 11.57    | 25.83    | 37.60    |

Numerous image-level contrastive learning methods have been proposed for object recognition and related tasks in recent years. However, it has not been exploited for learning correspondences especially on the semantic level. We show that image-level contrastive learning facilitates object-centric attention, which in turn helps finding semantic correspondence. Our approach is in direct contrast to existing semantic correspondence methods that mainly use ImageNet pretrained models for feature extraction.

Moreover, the proposed cross-instance pixel-level contrastive learning is specifically designed for semantic correspondence. Existing pixel-level contrastive learning approaches (Wang et al. 2020; Pinheiro et al. 2020; Jabri et al. 2020) are direct extensions of image-level instance contrastive learning (e.g., MoCo (He et al. 2020)). As such, they compute contrastive losses between different views of the same instance, e.g., either via self-augmented images (Wang et al. 2020; Pinheiro et al. 2020) or from sequences of the same instance (Jabri et al. 2020; Xie et al. 2021). Different from existing contrastive learning approaches, the proposed cross-instance pixel contrastive learning method leverages different images of the same category via cycle consistency. This is not a straightforward extension of the aforementioned approaches, especially when we do not leverage any pixel-level ground truth for constructing positive/negative sample as opposed to the scheme in (Kang et al. 2020). The proposed method aims to learn category-level mid-level representations, which is more challenging than the goals of existing approaches.

We further compare our method with other baselines on the SPair-71k dataset (Min et al. 2019b), which is a more challenging large-scale dataset. We use the self-supervised...
Fig. 4: Visualization of pixel-level contrastive learning. The first two columns show correspondence predictions across different images (i.e., the blue dots in the second column denote the correspondence of each pixel in $I'_0$). The remaining five columns show correspondence predictions from the augmented patch images $I'_0$ to their source images $I_0$, where the red dots denote the correspondence ground truth (where $I'_0$ is cropped from $I_0$), blue or green dots indicate the correspondence prediction based on the cycle affinity matrix $\bar{A}_{0'0} = A_{0'1}A_{10}$ or affinity matrix $A_{0'0}$. The last column shows the self-attention map on which the randomly cropped augmented image $I'_0$ is based. Note that the colored dots are computed based on the feature map. Thus the number of colored dots is less than the number of pixels.
Fig. 5: Visualization of the correspondence predictions on SPair-71k. The left and right columns are the predictions of our method and the ground truth, respectively. Different keypoints matches are indicated by different colors. The predictions are based on the voting matrix by going through the process of OT and RHM. See Section 4.4 for a detailed discussion.
correspondence.

posed method over the state-of-the-art schemes for semantic SPair-71k datasets demonstrate the effectiveness of the pro-

posed method in the same object in images or video sequences. Ex-
dense correspondence labels. This is different from existing the same category in different images without knowing their

correspondence. The pixel-level contrastive learning module con-
generating framework where the image-level contrastive learn-

ing framework from the validation set. In the pro-

out relying on supervised ImageNet pretrained models or

We pose a new task to learn semantic correspondence with-

more challenging than the PF-PASCAL and PF-WILLOW databases, as there exist large pose variations, viewpoint changes, scale differences, and occlusions in it. Fig. 5 presents some of those challenging cases where our method fails to obtain a good result. There are three directions to im-

prove our proposed method. First, self-supervised pretrained feature extractor needs to be improved to provide more robust feature representations by employing augmentations that can synthesize images with larger scale and viewpoint changes. Second, we may need to combine the cross cycle consistency regularization with a keypoint detector to learn more fine-grained correspondences. Third, the image-level contrastive learning could be improved by including more similar instances from different images to construct positive samples. Our future work will focus on addressing these issues.

5 Conclusion

We pose a new task to learn semantic correspondence with-

out relying on supervised ImageNet pretrained models or ground truth annotations from the validation set. In the pro-

posed method, we develop a multi-level contrastive learning framework where the image-level contrastive learning module generates object-level discriminative representations, and the pixel-level contrastive learning method con-

structs fine-grained representations to infer dense semantic correspondence. The pixel-level contrastive learning module is realized through the proposed cross-instance cycle consistency regularization, where we leverage different objects of the same category in different images without knowing their dense correspondence labels. This is different from existing approaches where correspondences are constructed by augmenting the same object in images or video sequences. Ex-
perimental results on the PF-PASCAL, PF-WILLOW, and SPair-71k datasets demonstrate the effectiveness of the proposed method over the state-of-the-art schemes for semantic correspondence.

Acknowledgements T. Xiao and M.-H. Yang are supported in part by NSF CAREER grant 1149783.

References

Bristow H, Valmadre J, Lucey S (2015) Dense semantic correspon-
dence where every pixel is a classifier. IEEE International Con-
ference on Computer Vision (ICCV) pp 4024–4031 1

Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple frame-
work for contrastive learning of visual representations. Interna-
tional Conference on Machine Learning (ICML) 2, 4, 5

Chen YC, Huang PH, Yu LY, Huang JB, Yang MH, Lin YY (2018) Deep semantic matching with foreground detection and cycle-

consistency. In: Asian Conference on Computer Vision (ACCV), Springer, pp 347–362 3, 4

Choy CB, Gwak J, Savarese S, Chandraker M (2016) Universal cor-

respondence network. Neural Information Processing Systems (NeurIPS) 1

Dalal N, Triggs B (2005) Histograms of oriented gradients for hu-

man detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 1

Dale K, Johnson MK, Sunkavalli K, Matusik W, Pfister H (2009) Image restoration using online photo collections. IEEE International Conference on Computer Vision (ICCV) pp 2217–2224 1

Deng J, Dong W, Socher R, Li Li, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 248–255 9

Doersch C, Gupta A, Efros AA (2015) Unsupervised visual representa-
tion learning by context prediction. IEEE International Conference on Computer Vision (ICCV) pp 1422–1430 4

Duchenne O, Joulia A, Ponce J (2011) A graph-matching kernel for object categorization. In: IEEE International Conference on Com-
puter Vision (ICCV), pp 1792–1799 1

Everingham M, Eslami SA, Van Gool L, Williams CK, Winn J, Zisserman A (2014) The pascal visual object classes challenge: A retrospective. International Journal on Computer Vision (IJCV) 111:98–136 7

Gidaris S, Singh P, Komodakis N (2018) Unsupervised representation learning by predicting image rotations. International Conference on Learning Representations (ICLR) 4

Grill JB, Strub F, Altchê F, Tallec C, Richemond PH, Bachschataya E, Doersch C, Pires BA, Guo ZD, Azar MG, et al. (2020) Bootstrap your own latent: A new approach to self-supervised learning. Neural Information Processing Systems (NeurIPS) 2, 4, 5

Ham B, Cho M, Schmid C, Ponce J (2016) Proposal flow. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 3475–3484 7

Ham B, Cho M, Schmid C, Ponce J (2018) Proposal flow: Semantic correspondences from object proposals. IEEE Transactions on Pat-
tern Recognition and Machine Intelligence (PAMI) 40:1711–1725 7

Han K, Rezende RS, Ham B, Wong KYK, Cho M, Schmid C, Ponce J (2017) Snet: Learning semantic correspondence. In: IEEE Interna-
tional Conference on Computer Vision (ICCV) 1, 2, 3, 4

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for im-
age recognition. IEEE Conference on Computer Vision and Pat-
tern Recognition (CVPR) pp 770–778 7

He K, Fan H, Wu Y, Xie S, Girshick R (2020) Momentum contrast for unsupervised visual representation learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 9729–9738 2, 3, 4, 5, 7, 9, 12, 15

Hjelm RD, Fedorov A, Lavioie-Marchildon S, Grewal K, Bachman P, Trischler A, Bengio Y (2019) Learning deep representations by mutual image information estimation and maximization. International Conference on Learning Representations (ICLR) 2, 4

pretrained network from MoCo (He et al. 2020) as initialization for all methods. As shown in Table 8, our method performs favorably against the HPF (Min et al. 2019a), DHPF (Min et al. 2020) and SCOT (Liu et al. 2020) approaches.

However, the results from the proposed self-supervised models are slightly worse than those by the supervised meth-

ods. To better understand the algorithmic performance, we use the self-supervised pretrained model as initialization and utilize some labeled image pairs with pixel correspondence ground truth for training. As shown in Table 9, leveraging pixel correspondence ground truth can improve model performance.
Xiao T, Hong J, Ma J (2018a) Dna-gan: Learning disentangled representations from multi-attribute images. International Conference on Learning Representations Workshop (ICLRW) 4
Xiao T, Hong J, Ma J (2018b) Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In: European Conference on Computer Vision (ECCV), pp 172–187 4
Xie Z, Lin Y, Zhang Z, Cao Y, Lin S, Hu H (2021) Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 16684–16693 4, 12
Yang H, Lin WY, Lu J (2014) Daisy filter flow: A generalized discrete approach to dense correspondences. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 3406–3413 1
Zhang R, Isola P, Efros AA (2016) Colorful image colorization. In: European Conference on Computer Vision (ECCV) 4
Zhou S, Xiao T, Yang Y, Feng D, He Q, He W (2017) Genegan: Learning object transfiguration and attribute subspace from unpaired data. In: British Machine Vision Conference (BMVC) 4
Zhou T, Lee YJ, Yu SX, Efros AA (2015a) Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 1191–1200 1, 2, 3
Zhou T, Krähenbühl P, Aubry M, Huang Q, Efros AA (2016) Learning dense correspondence via 3d-guided cycle consistency. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 117–126 1, 2, 3, 4
Zhu X, Zhu M, Daniilidis K (2015b) Multi-image matching via fast alternating minimization. In: IEEE International Conference on Computer Vision (ICCV), pp 4032–4040 4
Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision (ICCV), pp 2223–2232 4