Learning Semantic Alignment Using Global Features and Multi-Scale Confidence

Huaiyuan Xu, *Member, IEEE*, Jing Liao, *Member, IEEE*, Huaping Liu, *Senior Member, IEEE*, and Yuxiang Sun, *Member, IEEE*

***Abstract*— Semantic alignment aims to establish pixel correspondences between images based on semantic consistency. It can serve as a fundamental component for various downstream computer vision tasks, such as style transfer and exemplar-based colorization, etc. Many existing methods use local features and their cosine similarities to infer semantic alignment. However, they struggle with significant intra-class variation of objects, such as appearance, size, etc. In other words, contents with the same semantics tend to be significantly different in vision. To address this issue, we propose a novel deep neural network of which the core lies in global feature enhancement and adaptive multi-scale inference. Specifically, two modules are proposed: an enhancement transformer for enhancing semantic features with global awareness; a probabilistic correlation module for adaptively fusing multi-scale information based on the learned confidence scores. We use the unified network architecture to achieve two types of semantic alignment, namely, cross-object semantic alignment and cross-domain semantic alignment. Experimental results demonstrate that our method achieves competitive performance on five standard cross-object semantic alignment benchmarks, and outperforms the state of the arts in cross-domain semantic alignment.**

***Index Terms*— Semantic alignment, enhancement transformer, probabilistic correlation computation, cross-domain alignment.**

I. Introduction

IMAGE alignment is a fundamental computer vision task that identifies and corresponds the same/similar content between two images. Recently, taking advantage of the rich semantic features extracted by convolutional neural networks (CNNs), image alignment tasks are no longer limited to low-level visual scenarios, such as stereo matching [1], [2] and optical flow [3], [4], but can establish dense alignments between images based on high-level consistency of semantics. For example, aligning two cats that have obvious appearance variations but the same semantic contents (see Fig. 1). Semantic alignment, as a building block, facilitates many computer vision applications including style transfer [5], image translation/super-resolution [6], [7], and exemplar-based colorization [8].

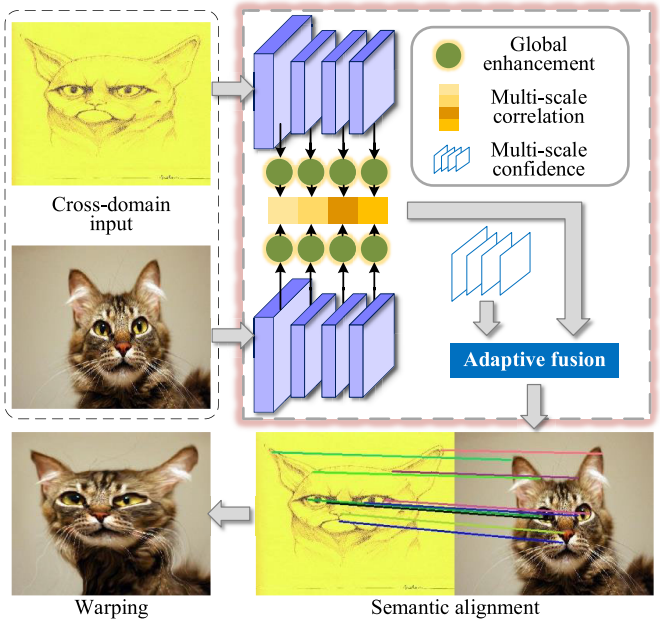


Fig. 1. The figures shows how we align two input images describing the same class but different instances according to semantic consistency. Accurate alignment can be obtained by global feature enhancement and adaptive multi-scale fusion. The colored lines visualize the keypoint alignment across the two bottom-right images. The bottom-left image shows the warped image based on the estimated semantic alignment.

The paradigm of most existing semantic alignment networks is as follows [9], [10], [11], [12], [13], [14], [15], [16], [17]: the CNN features are first extracted from two images, and then the correlations of feature point pairs are calculated to further guide the alignment decision. In this paradigm, there are two critical issues: one is semantic feature extraction; the other is accurate correlation estimation to reflect the similarity between points.

For the first issue, most methods use the pre-trained CNNs on the large-scale ImageNet dataset [18] to extract features [9], [10]. Although these CNN features contain coarse semantics, satisfactory for image-level classification, they are

Manuscript received 7 March 2023; accepted 6 June 2023. Date of publication 21 June 2023; date of current version 6 February 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62003286, in part by the Zhejiang Lab under Grant 2021NL0AB01, in part by the CCF-Baidu Open Fund under Grant 182215PCK04183, and in part by the HK PolyU under Grant P0034801. This article was recommended by Associate Editor Q. Wang. *(Corresponding author: Yuxiang Sun.)*

Huaiyuan Xu and Yuxiang Sun are with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: huaiyuan.xu@polyu.edu.hk; sun.yuxiang@outlook.com).

Jing Liao is with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong (e-mail: jingliao@cityu.edu.hk).

Huaping Liu is with the Department of Computer Science and Technology, Institute for Artificial Intelligence, Tsinghua University, Beijing 100084, China (e-mail: hpliu@tsinghua.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3288370.

Digital Object Identifier 10.1109/TCSVT.2023.3288370

1051-8215 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on February 07,2024 at 02:17:05 UTC from IEEE Xplore. Restrictions apply.

insufficient to describe fine-grained semantics for pixel-level matching between images. Therefore, it is necessary to further enhance the features. Some local operators [11], [14], [17] have been proposed to establish the neighborhood relevance of features and get more useful information from neighbors. However, this local enhancement pattern ignores global information that can benefit enhancement from a larger receptive field.

For the second issue, semantic correlations evaluate the semantic similarities of point pairs, serving as the direct cues for alignment assignment. In order to guarantee the accuracy of correlations, one idea is to learn convolution kernels to convolve the correlation map [14], [15], according to the space continuity property. Another idea is fusing multi-scale correlation maps [17] to produce better results. However, [17] directly sums correlation components without considering their confidence, so incorrect correlation components might mislead the network to output wrong alignments.

In this paper, to alleviate these problems, we propose the enhancement transformer and the probabilistic correlation module. The novelty of the enhancement transformer is that it can perceive global context to enhance semantic features. For example, given a human face image, if the face is globally perceived, it would help to identify the human eye. As for the probabilistic correlation module, it achieves adaptive correlation fusion with confidence estimation. According to estimated confidence probabilities, correlation components with low confidence scores are assigned with small fusion weights, thus their negative impact is reduced.

Moreover, we use the proposed network[1] to realize two alignment tasks, namely, cross-object semantic alignment and cross-domain semantic alignment [19]. The former focuses on images belonging to the same domain, such as a pair of photos [13], [14], [15], while the latter is different, concentrating on images with different domains, like an artwork and a photo. We propose a novel training strategy for cross-domain semantic alignment. It splits cross-domain semantic alignment into two sub-tasks, and then trains them simultaneously with weak supervision signals, which can learn robustness to domain variations and avoid tedious ground-truth alignment labeling. Extensive evaluations on multiple datasets demonstrate that our approach is comparable to the state-of-the-art algorithms in cross-object semantic alignment, and builds a state of the art in cross-domain semantic alignment. Ablation studies verify the effectiveness of our proposed components and training strategy. Furthermore, three applications are explored using our proposed method. Our contributions can be summarized as follows:

- We introduce a new enhancement transformer, which embeds global information into feature representations, enhancing features from a global perspective.
- We design a probabilistic correlation module, which measures multi-scale confidence scores to adaptively aggregate correlation components.
- We propose a novel training strategy for cross-domain semantic alignment, which alleviates the impact of domain variations and the problem of lacking alignment labels.

The remainder of this paper is organized as follows. Section II reviews related work. Section III introduces our approach, including the network architecture and training strategy. In Section IV, we analyze the results of quantitative, qualitative, and ablation experiments. Finally, we conclude this article and discuss future works in the last section.

II. Related Work

A. Semantic Features

The rapid development of semantic alignment benefits from the powerful semantic description capability of CNNs [20], [21] and the massive data in the ImageNet database [18]. Long et al. [20] first introduced CNN features into semantic alignment. However, the CNN features at that time did not have strong semantic invariance to large variances of appearance and shape. Afterward, some effective CNNs (e.g., VGGNet and ResNet pre-trained on ImageNet) are used to extract better semantic features. To further enhance these features, local self-similarity/-attention operators [11], [14], [17] are proposed to empower features to understand local context. However, objects in real images often have varying scales and shapes. Capturing semantic perception from the local context is not sufficient. In contrast, our method enhances features from a global perspective, that is, enabling them to perceive global information.

B. Semantic Correlation

The correlation map [9] is generated from the extracted semantic features, containing the matching scores for all possible matches between the two images. Then, the nearest neighbor can be obtained via retrieving the best match with the highest score. So, the correlation must effectively represent the degree of semantic consistency, otherwise, the wrong match would be retrieved. In other words, to improve alignment accuracy, it is necessary to enhance the original correlation. NCNet [22] first provided the idea by learning neighborhood consensus from the correlation map to enhance the original correlation. Following it, some variants [14], [15] appeared. Lee et al. [15] improves the efficiency of establishing neighborhood consensus through PatchMatch iteration [23]. Li et al. [14] designs non-isotropic 4-D convolution to adaptively explore neighborhoods. Another enhancement way is learning correlation complements from different scales, proposed by MMNet [17]. It assumes that all learned complements are valid and thereby adds them up. But this assumption does not hold in practice, as the complements obtained from different scales have different importance to semantic alignment, besides, there might be some incorrect complements. In contrast, our method computes the confidence probabilities of multi-scale correlations, as weights of correlation aggregation for better enhancement.

C. Training of Semantic Alignment Networks

Semantic alignment can be divided into cross-object semantic alignment and cross-domain semantic alignment

[1]Our code is available at: https://github.com/lab-sun/LearningSA

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on February 07,2024 at 02:17:05 UTC from IEEE Xplore. Restrictions apply.

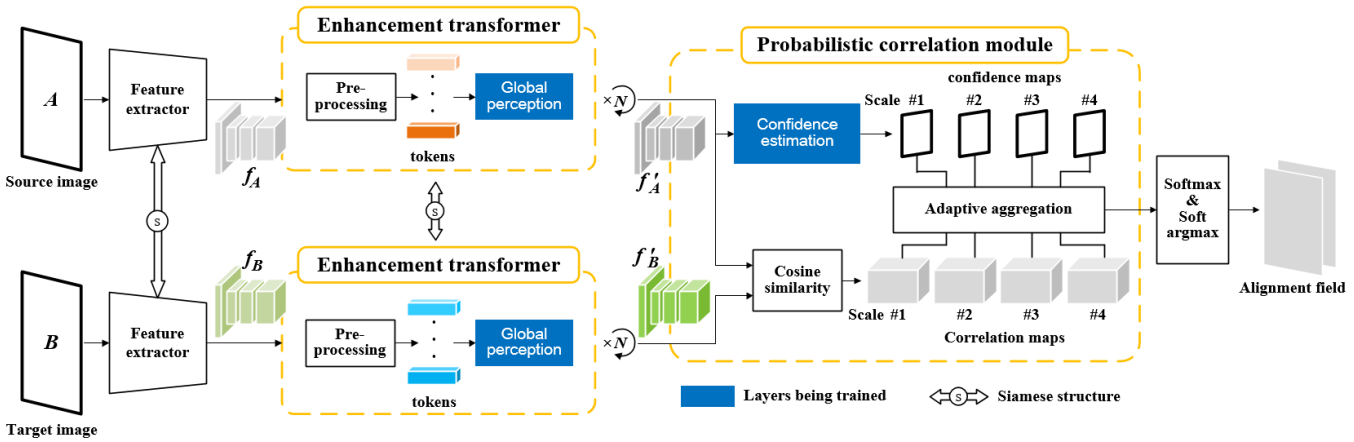


Fig. 2. The pipeline of the proposed framework. Given an input image pair, a pre-trained CNN extracts its feature maps (f_A, f_B). The feature maps then undergo two key components of our method: the enhancement transformer that globally enhances input features; the probabilistic correlation module that estimates multi-scale confidence maps for adaptive correlation aggregation. Finally, the alignment field is obtained by applying Softmax and Soft argmax to the output of the probabilistic correlation module.

(see details in Sec. I). In order for the network to learn the above two alignments, different training strategies are required. For learning the cross-object semantic alignment, one way is to train the network with constructed ground truth. For example, using known 3-D models to render synthetic images [21] of which ground-truth synthetic-to-synthetic correspondences can be computed; synthesizing image pairs [9], [10] by assumed transformations where per-pixel alignments are known; constructing positive and negative image pairs as the ground truth of weak supervision [11], [22] then maximizing the matching score of positive pairs. Another way is utilizing the ground truth, like keypoints [17] and masks [24], directly provided by existing datasets to train the networks. For learning the cross-domain semantic alignment, there is no labeled semantic alignments, thereby training in a strong supervised manner is impossible. Combining it with other tasks for weakly-supervised training is an alternative, for example, [19] uses the ground truth of image translation to train both translation and alignment tasks. We also propose a training strategy with weak supervision, but differently, we train the two semantic alignment sub-tasks simultaneously.

D. Transformer for Vision

Transformer is proposed by Vaswani et al. [25], and has revolutionized machine translation and natural language processing [26], [27]. Transformer has the advantage of global attention which can capture long-range relevancy. Recently, it has been successfully applied to diverse vision tasks, such as object detection [28], semantic segmentation [29], image classification [30], image captioning [31], etc. The drawback of transformer is the huge amount of parameters, resulting in large-memory consumption. The computation complexity of multi-head self-attention in the original transformer [25] is proportional to the square of the input size. In our approach, we design a lightweight transformer for the semantic alignment task and perform semantic enhancement on the relatively small-sized feature map.

III. The Proposed Approach

In this section, we describe our semantic alignment neural network and present the training details. In section III-A, two novel components for improving semantic alignment accuracy are discussed. One is the enhancement transformer which attends global context. The other is a probabilistic correlation module that learns multi-scale alignment confidence. In section III-B, we describe two training strategies, encouraging the network to learn semantic alignments in cross-object and cross-domain scenarios, respectively.

A. Network Architecture

The whole network architecture is illustrated in Fig. 2. Given as input an image pair (A, B), the network first uses a pre-trained feature extractor [32] to obtain deep feature maps (f_A, f_B) from input images. Then, the enhancement transformers provide global awareness to (f_A, f_B), and an probability correlation module outputs the adaptive aggregated correlation map for matching assignment.

1) Enhancement Transformer: The feature maps extracted by the pre-trained network [32] contain coarse semantics, which is satisfactory for image classification. However, for semantic alignment, it is necessary to further enhance the features to represent fine-grained semantics for pixel-level matching between images. Intuitively, the enhancement can be achieved by establishing the neighborhood relevancy of features and getting more useful information from neighbors [11], [14], [17]. But this enhancement ignores global information that can provide more representation enhancement. For example, an eye can be recognized from local perception, while combining with global information, it can further determine that this might be a left eye of human beings.

We cast the enhancement problem as a transformer to enhance features from a global perspective. The proposed enhancement transformer is shown in Fig. 3, which can be roughly summarized as that the input feature map is pre-processed to be a series of tokens, then they go through the global perception block to obtain global awareness.

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on February 07,2024 at 02:17:05 UTC from IEEE Xplore. Restrictions apply.

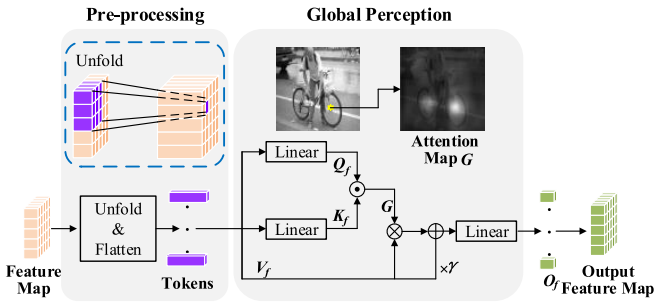


Fig. 3. The architecture of the enhancement transformer. It consists of two parts: the first part is the pre-processing to tokenize the feature map; the second part is the global perception block to provide global awareness to features. (Q_f, K_f, V_f) are three basic elements (query, key, value) of the attention mechanism [25]. O_f represents the output tokens and γ is a learnable parameter.

Pre-processing the feature map via the unfolding and flattening operation can tokenize the feature map, and help each token gather neighborhood information. To be specific, for the input feature map $f_{in} \in \mathbb{R}^{d_f \times h_f \times w_f}$, we concatenate the features in each $n \times n$ neighborhood to generate an unfolded feature map $f_{uf} \in \mathbb{R}^{n^2 d_f \times h_f \times w_f}$, then flatten it to $h_f w_f$ tokens by re-organizing it to $f_{tk} \in \mathbb{R}^{n^2 d_f \times h_f w_f}$. They are subsequently fed to a global perception block.

The global perception block explores long-range relevance by learning a global attention map. We first define query $Q_f = \text{Linear}(f_{tk})$, key $K_f = \text{Linear}(f_{tk})$, and value $V_f = f_{tk}$, where $\text{Linear}(\cdot)$ is the linear projection function to embed tokens. V_f is equal to the input to retain the original information extracted by the pre-trained network. Second, an attention map $\mathcal{G}_f$ is computed to describe the global relevance between tokens, by the dot product of the query and key, divided by $n\sqrt{d_f}$, and then followed by a Softmax function. Under the guidance of this attention map, we can obtain globally enhanced tokens O_f by the weighted summation of the value V_f and its global complement $\mathcal{G}_f \cdot V_f$, followed with a linear projection:

$$O_f = \text{Linear}\left((1-\gamma) \cdot \mathcal{G}_f \cdot V_f + \gamma \cdot V_f\right), \tag{1}$$

where $\gamma \in (0, 1)$ is a learnable factor. Notably, we adopt a sharing mechanism to share the parameters of all $\text{Linear}(\cdot)$ functions to lighten the enhancement transformer, in which only one linear projection and one weight factor need to be learned. Finally, we reshape O_f to be $f_{out} \in \mathbb{R}^{d_f \times h_f \times w_f}$, making the input and output sizes consistent in the enhancement transformer.

The computational complexity of the proposed enhancement transformer depends on three parts, namely linear projection, attention map, and weighted summation in Eq. 1. The linear projection maps $n^2 d_f$-dimensional features to d_f-dimensional ones, and its computational complexity is $\mathcal{O}(h_f w_f d_f^2 n^2)$. The computational complexity $\mathcal{O}(h_f^2 w_f^2 d_f)$ of the attention map comes from computing self-attention among features. The weighted summation has a computational complexity of $\mathcal{O}(h_f^2 w_f^2 d_f n^2)$, which comes from matrix multiplication $\mathcal{G}_f \cdot V_f$. As a result, the computational complexity of the enhancement transformer can be obtained as $\mathcal{O}(h_f w_f d_f^2 n^2 + h_f^2 w_f^2 d_f n^2)$, related to the feature map size (h_f, w_f, d_f) and the neighborhood range n.

2) Probabilistic Correlation Module: The correlation map stores the scalar products of point pairs between two feature maps, representing point-pair alignment scores. Fusing multi-scale correlation maps is proven to boost the alignment accuracy in [17], which directly adds multi-scale correlations to achieve fusion. However, the direct summation of multi-scale correlations would enlarge the effect of low-contribution correlations, while reducing the importance of high-contribution correlations. To alleviate this problem, we introduce the probabilistic correlation module, as illustrated in Fig. 4. It learns multi-scale confidence maps, then guides multi-scale correlation maps for adaptive aggregation.

On the one hand, a global weight g_i is learned to reflect the contribution of the i-th scale. On the other hand, we introduce local weights to describe the spatial variation of confidence, that is, the confidence varies from point to point. Specifically, we compute the channel-wise accumulation of the feature map $f'_{A,i}$ as semantic activation scores. The more semantics are activated, the easier for a point to find its match. Then, the activation scores undergo normalization and a bias addition to adjust the values to an appropriate range. Finally, the confidence map Z_i of the i-th scale is calculated by multiplying the global weight and the local weights:

$$Z_i = g_i \cdot \left(\text{ø}\left(f'_{A,i}\right) + e\right), \tag{2}$$

where ø is the operation of channel-wise summation and min-max normalization; e is a regularization coefficient.

We aggregate multi-scale correlation maps using confidence maps according to the law of total probability. Let $P(m)$ be the marginal probability of the event where a point pair m is semantically aligned. It is equivalent to the summation of occurring probabilities of this event under different conditions, i.e., the weighted average of conditional probabilities $P(m|S_i, i = 1, \ldots, l)$ of that m is a semantic match under different scales S_i:

$$P(m) = \sum_{i=1}^{l} P(m \cap S_i) = \sum_{i=1}^{l} P(m|S_i)\, P(S_i), \tag{3}$$

where $P(S_i)$ represents the occurrence probability of the S_i condition, which can be sampled from the confidence map Z_i. $P(m|S_i)$ measures the alignment probability at the S_i condition, which can be obtained from the correlation map C_i of the i-th scale. Then, the multi-scale aggregated correlation map Λ can be computed as:

$$\Lambda(m) = \sum_{i=1}^{l} C_i(m)\, Z_i(m). \tag{4}$$

Here, the confidence maps are normalized before being used for aggregation, since they are required to form an entire sample space, that is $\sum_{i=1}^{l} Z_i(m) = 1$.

3) Matching Assignment: We assign each point with a semantic alignment from the correlation map Λ. Specifically, we compute the semantic mapping $p \rightarrow q$ of point p in

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on February 07,2024 at 02:17:05 UTC from IEEE Xplore. Restrictions apply.

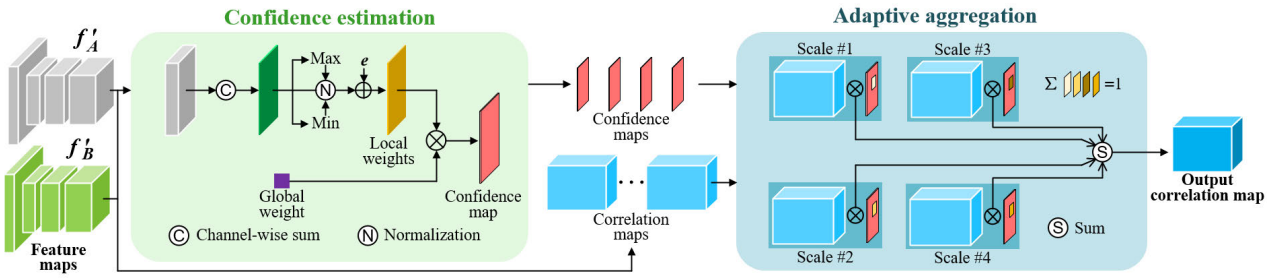


Fig. 4. The architecture of the probabilistic correlation module. When given input a pair of feature maps, it outputs an aggregated correlation map. The confidence estimation block learns multi-scale confidence maps, which are subsequently fed to the adaptive aggregation block to guide the fusion of the correlation maps from different scales.

the feature map f_A by calculating an average position of all candidates in the feature map f_B with correlations as weights:

$$p \to q = \sum_{q \in f_B} \text{softmax}\left(\beta \cdot \Lambda\left(p, q\right)\right) \cdot q. \tag{5}$$

Here, β is the coefficient that controls the sharpness of the Softmax function.

B. Training Strategies

Our semantic alignment network can establish semantic correspondences in different scenarios, such as estimating alignments across photos (cross-object) or between the artwork and photo (cross-domain). For cross-object alignment, some datasets provide sparse ground-truth keypoint annotations [33], [34], whose alignment relationships can be used as strong supervision signals for network training. Unfortunately, the cross-domain scenario has no ground-truth alignment labels for training. Thus, we propose a novel training strategy that only requires the object category label, such as two images of dogs, to achieve weakly-supervised learning.

1) Cross-Object Alignment Learning With Strong Supervision: Compared with labels of object categories [22] and foreground masks [24], keypoint labels can provide more concrete semantics, such as eyes and mouths. Point-wise matching between keypoints is directly linked to the semantic alignment task, so it can provide a strong supervision signal. Specifically, we use a multi-scale landmark loss to train the network by minimizing the Euclidean distance between ground-truth keypoint p in the source image and the estimated one p' by translating its corresponding target keypoint q back to the source with the predicted alignment:

$$\mathcal{L}_{land} = \frac{1}{N(l+1)} \sum_{i=1}^{l+1} \sum_{j=1}^{N} \left\| p_j - p'_{i,j} \right\|_2, \tag{6}$$

where j represents the j-th keypoint. i indicates the i-th scale. As for the scale of the aggregated correlation map, we denote it as the $l+1$ scale.

2) Cross-Domain Alignment Learning With Weak Supervision: Unlike the cross-object dataset, the cross-domain dataset does not have semantic alignment labels for training. Compared with using auxiliary supervision from other tasks [19], we propose a novel training strategy with more direct supervision by training two alignment sub-tasks. Specifically, we split cross-domain semantic matching into the same-object but cross-domain (SOCD) matching and the cross-object but same-domain (COSD) matching, as shown in Fig. 5. By training both sub-tasks simultaneously, the network learns insensitivity to domain variation and intra-class variation.

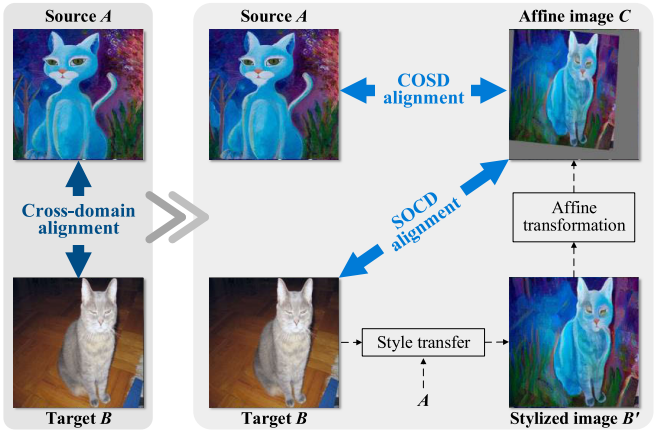


Fig. 5. Cross-domain alignment can be divided into cross-object but same-domain (COSD) matching, and same-object but cross-domain (SOCD) matching. We randomly select an artwork as source image A and a photo as target image B. To synthesize a transition image C, we transfer the style of A to B [35] then deform stylized B' by a random affine transformation.

For SOCD matching, since the affine image C (in Fig. 5) is generated from the stylized target image B' with a random but known affine transformation, the real dense mapping $M_{C \to B}$ from C to B is known. Therefore, we define a mapping loss to encourage the estimated mapping $\widehat{M}_{C \to B}$ to be consistent with the real one:

$$\mathcal{L}_{map} = \frac{1}{HW} \left\| \widehat{M}_{C \to B} - M_{C \to B} \right\|_2, \tag{7}$$

where HW is the size of the image.

For COSD matching, since image C has the same style as source image A, its warped image based on predicted alignment should be similar to A. Thus, we use pixel loss $\mathcal{L}_{px}$ and feature reconstruction loss $\mathcal{L}_{feat}$ [36] to encourage reconstruction of A from C in RGB and feature spaces, respectively. Furthermore, we employ forward-backward consistency loss $\mathcal{L}_{cons}$ [37] to encourage one-to-one mapping between A and C. The overall loss for training cross-domain alignment network can be interpreted as a weighted summation of the

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on February 07,2024 at 02:17:05 UTC from IEEE Xplore. Restrictions apply.

above losses:

$$\mathcal{L}_{c-dom} = \lambda_{map}\mathcal{L}_{map} + \lambda_{px}\mathcal{L}_{px} + \lambda_{feat}\mathcal{L}_{feat} + \lambda_{cons}\mathcal{L}_{cons}, \tag{8}$$

where $(\lambda_{map}, \lambda_{px}, \lambda_{feat}, \lambda_{cons})$ are the weighting parameters.

IV. Experimental Results and Discussions

In this section, we first describe the implementation details and the common evaluation metric for semantic alignment. We then perform quantitative and qualitative analyses to compare the performance of our approach against state-of-the-art methods. The ablation study verifies the effectiveness of the network architecture as well as the proposed training strategy. We also analyze memory cost, runtime, and limitations of our method. Finally, some interesting applications are explored.

A. Implementation Details

Our network is implemented using PyTorch and trained on a 24GB NVIDIA GeForce RTX 3090 GPU. In the network, the feature extractor uses pre-trained ResNet-101 [32] of which parameters are fixed during training. In the enhancement transformer, the unfolded area is set to 3×3 neighborhood. Two enhancement transformer layers are stacked for feature enhancement. In the probabilistic correlation module, the number of scales is set to 4, corresponding to 4 stages of the pre-trained ResNet. The coefficient β of the matching assignment part is set to 100. The loss coefficients (λ_{map}, λ_{rec}, λ_{perc}, λ_{cons}) of the training strategy are set to (1, 10, 10, 0.5). We use bilinear interpolation to upsample the output alignment field to 320×320, consistent with the size of input images. During training, we use the Adam optimizer for cross-object alignment learning with a learning rate 3×10^{-5} following [24]. The learning rate 2×10^{-4} for cross-domain alignment is determined by grid search.

B. Evaluation Metric

We compare our method with previous methods by the commonly-used metric: the probability of correct keypoint (PCK) [45]. This quantitative metric computes the percentage of keypoints whose alignment errors below $\alpha \cdot \max(h, w)$, where h and w represent the height and width of either an image (αimg) or an object bounding box (αbbox). Assuming that in the test set, $\mathcal{P}_k$ is a set of source and target keypoint pairs (p_s, p_t) of k-th image pair, then we have:

$$\text{PCK}_k = \frac{1}{|\mathcal{P}_k|} \sum_{(p_s, p_t) \in \mathcal{P}_k} \mathbb{1}\left[D\left(\mathcal{T}_k\left(p_s\right), p_t\right) < \alpha \cdot \max(h, w)\right], \tag{9}$$

where D calculates the Euclidean distance, $\mathcal{T}_k$ is the estimated semantic alignment field, and $\mathbb{1}$ is the indicator function that returns 1 if the expression inside is true and 0 otherwise. The final PCK score is the average of all PCK_k of test image pairs.

C. Quantitative Evaluation on Benchmarks

We compare our approach to the state-of-the-art methods of cross-object semantic alignment on five popular benchmarks: PF-PASCAL [33], PF-WILLOW [47], SPair-71K [34], Caltech-101 [48], and TSS [49]. However, there is no benchmark for CROss-DOMain semantic alignment, so we build a CroDom dataset containing artworks and photos to analyze the performance of our algorithm in the cross-domain scenario.

1) PF-PASCAL [33] and PF-WILLOW [47]: The PF-PASCAL benchmark contains 20 object categories with 6 to 140 image pairs in each category. In each image pair, it provides 4 to 17 keypoint annotations. We follow the training/valid/test splits of [14]. PF-WILLOW benchmark includes 4 object classes that are further subdivided into 10 subtypes. It provides a total of 900 image pairs, where each image has 10 keypoint annotations. These two benchmarks are more challenging than previous datasets [48], [49] because of larger variations of object appearance and scene layout. Following [38], we use the PCK metric for PF-PASCAL with αimg and PF-WILLOW with a more strict αbbox.

We test our method on PF-PASCAL and PF-WILLOW, respectively. When testing on PF-WILLOW, we directly use the network trained on PF-PASCAL without any fine-tuning. The quantitative comparison with the state of the arts is shown in Tab. I, and per-class performances on PF-PASCAL are present in Tab. II. For comparative methods, we take the alignment results of CATs [42] and TransforMatcher [44] without extra tricks, namely fine-tuning features and data augmentation respectively, to provide a fair comparison to reflect the performance of the network structure. It can be seen from the tables that: (1) Compared with other supervisions, using a small number of sparse landmarks as supervision signals to train the network can achieve higher PCK scores; (2) Our method achieves (81.3%, 92.9%) and (55.6%, 80.4%) PCK scores on PF-PASCAL and PF-WILLOW respectively, comparable to the state-of-the-art methods in alignment accuracy; (3) Our method obtains PCK scores over 90.0% on 14/20 object categories of PF-PASCAL, indicating robustness to object category changes; (4) Compared with the performance on PF-PASCAL, the PCK scores on PF-WILLOW decrease by (25.7%, 12.5%). This is because we do not retrain the network on the PF-WILLOW dataset. The result shows that our method has a competitive generalization ability to a novel dataset despite the data-distribution change.

2) SPair-71k [34]: This is a newly-released cross-object semantic alignment benchmark. It consists of 70, 958 image pairs of 18 object classes, including 12, 234 pairs for testing. It provides diverse variations in viewpoint, scale, truncation, and occlusion. Compared with other datasets, the most significant characteristic is the large scale and detailed splits of the data. The PCK threshold of SPair-71K is calculated using the size of bounding-box with αbbox $= 0.1$, consistent with the previous works [12], [38] for fair comparison.

Tab. III reports per-class semantic matching performance, involving animals, plants, vehicles, etc. In each class, images suffer from varying degrees of object size and perspective

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on February 07,2024 at 02:17:05 UTC from IEEE Xplore. Restrictions apply.

TABLE I

QUANTITATIVE EVALUATION ON PF-PASCAL, PF-WILLOW, CALTECH-101, AND TSS BENCHMARKS. THE BEST PERFORMANCE IS IN BOLD, AND THE UNDERSCORED ONE IS THE SECOND BEST. RESULTS (%) OF [9], [10], [11], [22], [24], AND [38] ON PF-PASCAL AND PF-WILLOW ARE BORROWED FROM [12]. CALTECH-101 AND TSS RESULTS ARE FROM [13] AND [39], RESPECTIVELY

Methods	Supervision	PF-PASCAL PCK@α_{img} (%)		Spair-71K PCK@α_{bbox} (%)	PF-WILLOW PCK@α_{bbox} (%)		Caltech-101		TSS PCK@α_{img}, $\alpha = 0.05$ (%)		
		0.05	0.10	0.10	0.05	0.10	LT-ACC	IoU	FG3D	JODS	PASCAL
CNNGeo [9]	Synthetic Warp	41.0	69.5	20.6	36.9	69.2	0.79	0.56	-	-	-
A2Net [10]		42.8	70.8	22.3	36.3	68.8	0.80	0.57	-	-	-
NCNet [22]	Class Labels	54.3	78.9	20.1	33.8	67.0	0.85	0.60	92.3	76.9	57.1
DUS [40]		-	-	34.0	-	-	-	-	-	-	-
DCCNet [11]		-	82.3	-	43.6	73.8	-	-	-	-	-
SFNet [24]	Masks	-	78.7	-	-	74.0	**0.88**	**0.67**	88.0	75.1	58.4
HPF [38]	Landmarks	63.5	88.3	28.2	48.6	76.3	**0.88**	0.64	93.6	79.7	57.3
ANCNet [14]		-	88.7	30.1	-	-	-	-	-	-	-
SCOT [12]		67.3	88.8	35.6	50.7	78.1	-	-	95.3	81.3	57.7
DHPF [13]		75.7	90.7	37.3	49.5	77.6	0.87	0.62	88.2	71.9	56.6
PMD [16]		-	90.7	37.4	-	75.6	-	-	-	-	-
CHM [41]		80.1	91.6	46.3	52.7	79.4	-	-	-	-	-
CATs† [42]		67.5	89.1	42.4	46.6	75.6	-	-	-	-	-
PMNC [15]		**82.4**	90.6	50.4	-	-	-	-	-	-	-
SemiMatch [43]		75.0	91.7	43.0	47.4	76.3	-	-	-	-	-
TransforMatcher [44]		78.9	90.5	50.2	-	75.1	-	-	-	-	-
PWarpC [39]		79.2	92.1	37.1	48.0	76.2	-	-	**97.5**	**87.8**	**88.4**
MMNet [17]		81.1	91.6	50.4	-	-	-	-	-	-	-
Ours	Landmarks	81.3	**92.9**	**51.1**	**55.6**	**80.4**	**0.88**	0.65	95.8	82.3	63.9

TABLE II

PER-CLASS AND AVERAGE ALIGNMENT ACCURACY ON THE PF-PASCAL DATASET. THE BEST RESULTS (%) ARE IN BOLD, AND THE SECOND BESTS ARE UNDERLINED. RESULTS OF [22] AND [46] COME FROM [11]. THE REST ARE FROM THE CORRESPONDING PAPERS OR SOURCE-CODE IMPLEMENTATIONS

Methods	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	d.table	dog	horse	moto	person	plant	sheep	sofa	train	tv	all
UCN-ST [46]	64.8	58.7	42.8	59.6	47.0	42.2	61.0	45.6	49.9	52.0	48.5	49.5	53.2	72.7	53.0	41.4	83.3	49.0	73.0	66.0	55.6
NCNet [22]	86.8	86.7	86.7	55.6	82.8	88.6	93.8	87.1	54.3	87.5	43.2	82.0	64.1	79.2	71.1	71.0	60.0	54.2	75.0	82.8	78.9
DCCNet [11]	87.3	88.6	82.0	66.7	84.4	89.6	94.0	90.5	64.4	91.7	51.6	84.2	74.3	83.5	72.5	72.9	60.0	68.3	81.8	81.1	82.3
SCOT [12]	88.8	94.7	87.7	**94.4**	90.6	96.9	97.2	91.4	70.4	91.7	75.0	90.5	84.3	89.2	81.4	90.0	80.0	76.2	91.0	81.7	88.8
DHPF [13]	95.6	91.4	84.7	81.9	85.9	94.5	96.0	91.5	79.7	**95.8**	79.7	**94.7**	88.0	91.0	90.5	94.8	**100**	85.7	87.0	**96.7**	90.7
CHM [41]	**96.2**	93.6	91.6	87.5	79.7	**99.2**	96.1	92.3	76.7	**95.8**	84.4	89.2	92.1	93.5	88.6	89.1	**100**	83.1	90.0	**96.7**	91.6
PWarpC [39]	93.3	**96.9**	88.9	81.9	**95.3**	**99.2**	**97.9**	92.3	77.8	93.8	81.2	94.1	86.6	93.4	89.0	94.3	**100**	83.8	88.0	91.7	92.1
MMNet [17]	93.1	91.8	**93.3**	79.2	**95.3**	**99.2**	97.8	**92.4**	**88.1**	87.5	73.4	90.2	**93.0**	**94.1**	**93.5**	89.5	**100**	83.8	88.0	**96.7**	91.6
Ours	93.8	96.1	90.0	91.7	93.8	**99.2**	**97.9**	89.7	86.7	93.8	**85.9**	84.7	**93.0**	92.6	85.7	**100**	**100**	**88.0**	**94.0**	95.6	**92.9**

TABLE III

PER-CLASS AND AVERAGE ALIGNMENT ACCURACY ON THE SPAIR-71K DATASET. THE BEST RESULTS (%) ARE IN BOLD AND THE SECOND BESTS ARE UNDERLINED. RESULTS OF [9], [10], [12], [13], [22], AND [38] COMES FROM [17]. THE REST RESULTS ARE BORROWED FROM THE CORRESPONDING PAPERS OR SUPPLEMENTARY MATERIALS

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	dog	horse	moto	person	plant	sheep	train	tv	all
CNNGeo [9]	23.4	16.7	40.2	14.3	36.4	27.7	26.0	32.7	12.7	27.4	22.8	13.7	20.9	21.0	17.5	10.2	30.8	34.1	20.6
A2Net [10]	22.6	18.5	42.0	16.4	37.9	30.8	26.5	35.6	13.3	29.6	24.3	16.0	21.6	22.8	20.5	13.5	31.4	36.5	22.3
NCNet [22]	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
HPF [38]	25.2	18.9	52.1	15.7	38.0	22.8	19.1	52.9	17.9	33.3	32.8	20.6	24.4	27.9	21.1	15.9	31.5	35.6	28.2
SCOT [12]	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
DHPF [13]	38.4	23.8	68.3	18.9	42.6	27.9	20.1	61.6	22.0	46.9	46.1	33.5	27.6	40.1	27.6	28.1	49.5	46.5	37.3
PMD [16]	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
PMNC [15]	54.1	35.9	**74.9**	36.5	42.1	48.8	40.0	**72.6**	21.1	67.6	**58.1**	50.5	40.1	**54.1**	43.3	35.7	74.5	59.9	50.4
SemiMatch [43]	47.8	29.0	70.6	24.0	44.5	37.6	29.8	65.2	17.2	54.7	52.8	47.1	35.2	37.6	29.9	32.7	68.5	49.4	43.0
TransforMatcher [44]	54.5	33.9	72.2	**38.5**	47.7	55.3	45.6	65.7	25.2	62.6	58.0	47.0	40.7	44.2	43.1	35.3	71.9	61.6	50.2
MMNet [17]	**55.9**	37.0	65.0	35.4	50.0	**63.9**	**45.7**	62.8	**28.7**	65.0	54.7	51.6	38.5	34.6	41.7	**36.3**	**77.7**	62.5	50.4
Ours	54.4	**38.3**	72.8	30.4	**52.1**	46.3	36.6	68.0	28.4	**68.5**	52.3	**53.6**	**42.2**	36.3	**53.9**	31.2	76.1	**72.7**	**51.1**

differences, as well as potential occlusion and truncation. Tab. III shows that our method achieves the best performance in a total of 7 object categories, outperforming the second-ranked algorithm [17] that obtains the best in 6 categories. On the other hand, our method achieves an average PCK score of 51.1% in all classes, superior to other algorithms, demonstrating the robustness of our method to various object classes.

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on February 07,2024 at 02:17:05 UTC from IEEE Xplore. Restrictions apply.

Fig. 6. Image warping on the SPair-71K dataset. We utilize the predicted keypoint semantic alignment and the thin-plate spline transformation to warp A to A'. Ideally, the same locations in A' and B have the same semantics.

3) Caltech-101 [48] and TSS [49]: The Caltech-101 benchmark provides 1,515 image pairs from 101 categories with segmentation annotations. We transfer the ground-truth labels via predicted semantic alignment, and count the proportion of correctly transferred labels as the label transfer accuracy (LT-ACC) [50]. Besides, we adopt intersection-over-union (IoU) [51] to evaluate the quality of foreground label transfer. The TSS benchmark contains 400 image pairs annotated with dense flow fields, divided into 3 groups, namely FG3D, JODS, and PASCAL, according to the data sources. Following [12], we use PCK scores to evaluate alignment accuracy on TSS.

Similar to the PF-WILLOW benchmark, we train the network on PF-PASCAL and then test it on Caltech-101 and TSS. Quantitative results are shown in Tab. I. We can see that our method has competitive alignment accuracy compared with other state-of-the-art methods, indicating a comparable generalizability across different benchmarks. PWarpC [39] outperforms other methods on TSS. One possible explanation is that PWarpC has warp consistency constraints in the loss, which can provide additional constraints from non-landmark pixels. We leave this issue of loss for future study, which might further improve our performance on the TSS benchmark.

4) CroDom: To the best of our knowledge, we build the first cross-domain dataset called CroDom for semantic alignment. The images of CroDom are divided into two domains, namely photos and artworks, and they come from ImageNet [18] and BAM-dataset [52], respectively. Our collector collected more than 1,400 images and split them into 6 categories, that is, bicycle, bird, car, cat, dog, and person, according to known object-class labels. Fig. 7 presents some samples. 791 images are randomly selected to form the training set. The rest images are combined into 155/154 cross-domain image pairs as the validation/test set. Since the collected images have no pre-existing keypoint labels, our two labelers manually annotated 20-30 keypoint matching labels for each image pair in the validation and test sets, and checked them by visualization. These keypoint labels can be used to evaluate the cross-domain alignment performance of the network. Similar to the PF-PASCAL dataset, CroDom also adopts PCK scores as the quantitative evaluation metric.

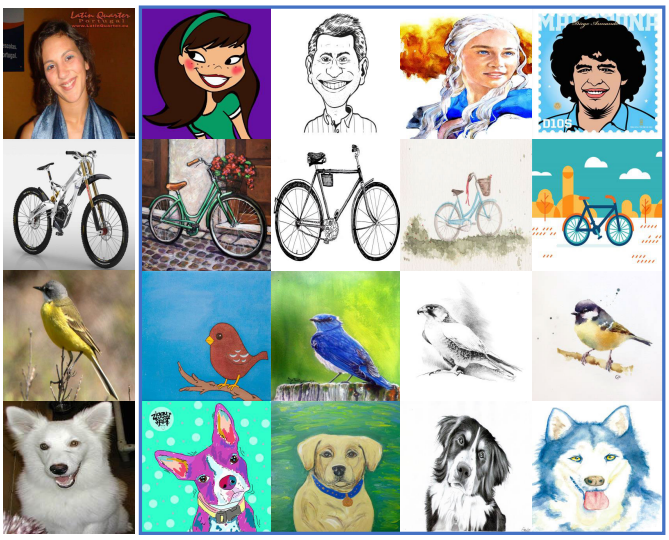


Fig. 7. Samples of the CroDom dataset. CroDom provides photos and artworks from the ImageNet database [18] and BAM dataset [52] respectively. Columns 2-5 display the artworks across various media, including comic, oil paint, pen ink, pencil sketch, watercolor, and vector art.

Cross-domain semantic alignment is a new branch of semantic matching. CoCosNet [19] is the state-of-the-art method, which utilizes the image generation task to assist the training of cross-domain alignment. We train CoCosNet on the CroDom training set by feeding it cross-domain image pairs consisting of photos and artworks. Using the style transfer method [35] can obtain the image generation ground-truth, as the supervision signal for training. Differently, our network is trained with image triplets, built using cross-domain image

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on February 07,2024 at 02:17:05 UTC from IEEE Xplore. Restrictions apply.

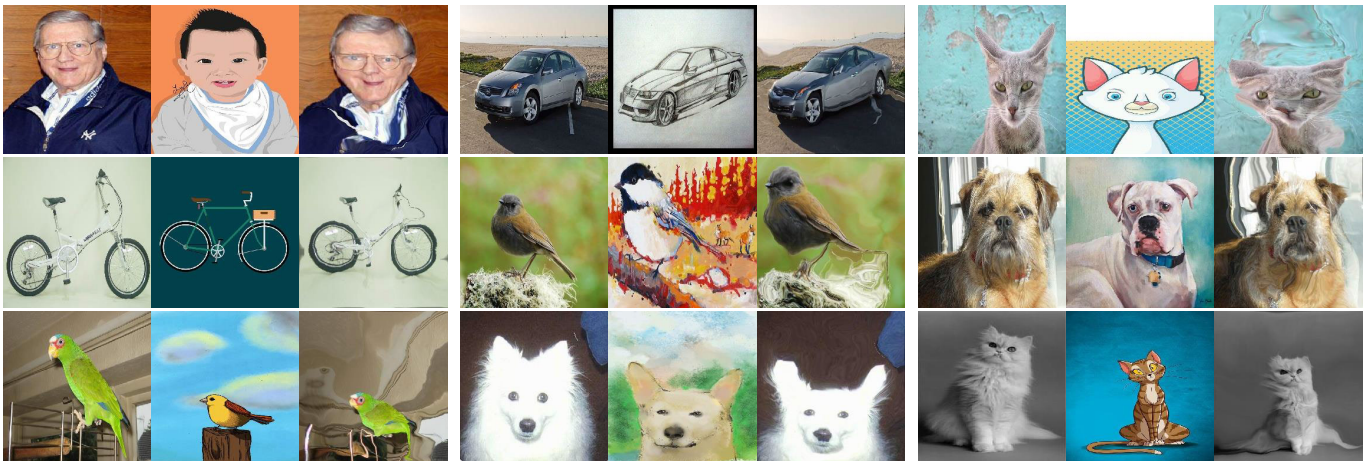


Fig. 8. Image warping on the CroDom dataset. We map A to A' pixel by pixel according to the semantic alignment between A and B. Ideally, the same positions in A' and B have the same semantics.

TABLE IV

QUANTITATIVE EVALUATION ON THE CRODOM DATASET. PCK SCORES (%) ARE REPORTED. THE BEST PERFORMANCES ARE IN BOLD. $*$ MEANS TO TRAIN THE NETWORK USING OUR PROPOSED TRAINING STRATEGY

Methods	PCK@α_{img}				
	$\alpha = 0.01$	$\alpha = 0.03$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$
CoCosNet [19]	3.7	25.6	48.1	77.6	89.2
MMNet$*$ [17]	4.7	29.3	50.9	79.2	88.9
Ours	**7.1**	**38.2**	**61.0**	**83.9**	**91.9**

TABLE V

PER-CLASS EVALUATION ON THE CRODOM DATASET. PCK SCORES (%) W.R.T. IMAGE SIZE WITH $\alpha = 0.10$ ARE REPORTED. THE BEST PERFORMANCES ARE IN BOLD. $*$ MEANS TO TRAIN THE NETWORK USING OUR PROPOSED TRAINING STRATEGY

Methods	bicycle	bird	car	cat	dog	person	all
CoCosNet [19]	85.5	81.0	87.4	74.7	66.7	81.8	77.6
MMNet$*$ [17]	79.5	84.2	85.5	78.8	70.5	81.2	79.2
Ours	**86.2**	**85.6**	**92.3**	**85.3**	**74.3**	**84.3**	**83.9**

pairs and style transfer, as shown in Fig. 5. Our network learns two alignment sub-tasks to achieve the learning of cross-domain semantic alignment. In addition, we select MMNet, one of the best methods on the SPair-71K dataset, and replace its original training strategy with our proposed cross-domain training strategy, so that it can achieve cross-domain semantic alignment without the need of landmark labels. Tab. IV and V show quantitative comparisons. It can be seen that our method outperforms CoCosNet and MMNet in average and per-class alignment accuracy, and sets up a new state of the art.

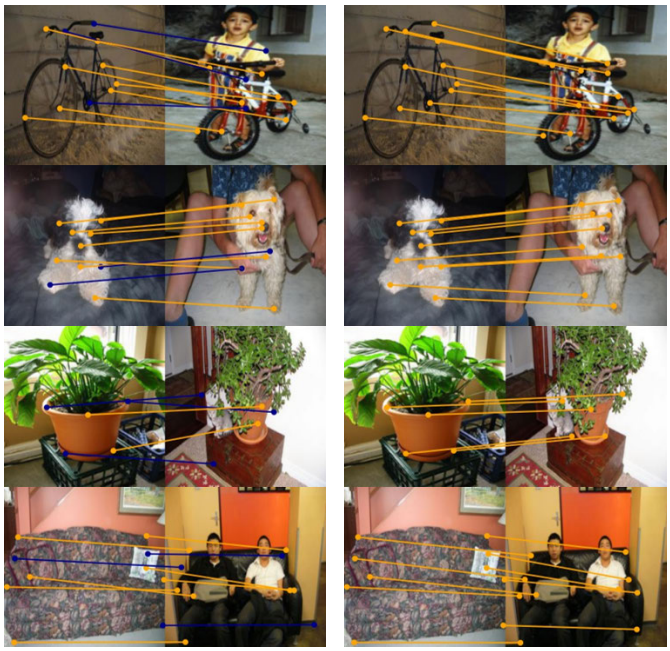
D. Qualitative Analysis

To qualitatively evaluate the robustness of our method, we perform semantic alignment-based image warping in different scenarios. Additionally, we make alignment visualization for the qualitative comparison with the baseline algorithm MMNet [17].

1) Image Warping: We select image pairs of different classes for semantic image warping. These classes include airplanes, boats, buses, cars, birds, cats, etc. Fig. 6 and Fig. 8 show the warping quality on the SPair-71K and CroDom datasets, respectively. Ideally, according to the estimated semantic alignment field, the warped image is semantically aligned with the target image, that is, the warped image and the target image have the same semantic content at the same position in the images. On SPair-71K, we utilize the estimated keypoint semantic alignment and the thin-plate spline interpolation to compute a dense semantic alignment field, which is further used to warp the image. As can be seen from Fig. 6, our method is insensitive to object scale changes (row 2, columns 1 to 3), background interference, obstacle occlusion (row 3, columns 4 to 6), object truncation (row 3, columns 4 to 6), and viewpoint variations. Notably, the tolerance for viewpoint change can reach 180 degrees (see row 4, columns 1 to 3). On CroDom, we implement image warping by pixel mapping based on the dense semantic alignment field output by our network. Fig. 8 shows that our method achieves accurate and smooth image warping between a photo and an artwork of different styles, such as cartoon, oil painting, pen drawing, and watercolor. In summary, our method can handle the variations of object category, image domain, viewpoint, object scale, etc., indicating the robustness under different input scene settings.

2) Qualitative Comparison With MMNet: We choose MMNet [17] as the baseline method, which enhances features with the local pattern and directly sums multi-scale correlations to improve semantic alignment accuracy. In contrast, our method enhances features from a global perspective and aggregates multi-scale correlations adaptively. We visualize the predictions of semantic matches to qualitatively compare our method with the baseline. In Fig. 9, the orange lines represent the correct estimates of semantic alignment, while the blue ones highlight incorrect estimates with alignment errors greater than $0.1 \cdot \max(h, w)$, where (h, w) are the height and width of the image. Fig. 9 shows that the results of our method have fewer false semantic alignments than the

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on February 07,2024 at 02:17:05 UTC from IEEE Xplore. Restrictions apply.



MMNet [17] Ours

Fig. 9. Comparison with the baseline MMNet [17]. Orange lines denote correct semantic alignment predictions, while blue lines represent wrong predictions with obvious alignment errors.

TABLE VI

ABLATION RESULTS (%) OF THE ENHANCEMENT TRANSFORMER. THE ENHANCEMENT TRANSFORMER, DENOTED BY 'ET', CAN BE REPLACED BY LOCAL SELF-SIMILARITY [14], CONTEXT ENCODER [11], LOCAL SELF-ATTENTION [17]. HOWEVER, THE ENHANCEMENT TRANSFORMER HAS HIGHER PCK SCORES THAN THESE LOCAL METHODS, INDICATING BETTER PERFORMANCE

Algorithm Variants	PF-PASCAL		PF-WILLOW	
	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$
w/o ET	18.1	43.3	15.0	39.4
Local Self-similarity [14]	22.9	49.7	19.9	40.5
Context Encoder [11]	53.8	75.3	34.0	58.5
Local Self-attention [17]	66.0	85.2	42.8	67.7
Ours	**81.3**	**92.9**	**55.6**	**80.4**

baseline method when the input images contain significant visual variations.

E. Ablation Study

In this section, we present ablation studies on the proposed architectural components and training strategy, so as to analyze their effectiveness.

1) Ablation on Enhancement Transformer: To verify the effectiveness of the proposed enhancement transformer with global awareness, we report quantitative evaluations on the PF-PASCAL and PF-WIILOW datasets when removing the enhancement transformer or replacing it with other local techniques, as shown in Tab. VI. The visualized alignment errors of keypoints are presented in Fig. 10.

As we can see, the enhancement transformer and other local methods can improve semantic alignment accuracy. However, using local self-similarity [14], context encoder [11] or local self-attention [17] to substitute our enhancement transformer would cause PCK reduction, indicating the decrease of semantic alignment accuracy. This performance degradation is due to the lack of global perception, as the local techniques only attend the neighborhood features without considering the long-range relevance of features like our enhancement transformer. Fig. 10 qualitatively demonstrates that the enhancement transformer can effectively reduce the semantic matching ambiguity of keypoints.

TABLE VII

THE ABLATION RESULTS (%) OF THE PROBABILISTIC CORRELATION MODULE. THE PROBABILISTIC CORRELATION MODULE IS DENOTED AS 'PCM'. WHEN WE REMOVE THE PCM OR ITS KEY BLOCKS/COMPONENTS, SEMANTIC ALIGNMENT ACCURACY DECREASES WITH LOWER PCK SCORES, INDICATING THEIR IMPORTANCE TO SEMANTIC ALIGNMENT

Algorithm Variants	PF-PASCAL			
	$\alpha = 0.01$	$\alpha = 0.03$	$\alpha = 0.05$	$\alpha = 0.10$
w/o PCM	13.4	50.1	71.3	90.7
w/o Confidence Est.	23.5	64.4	79.7	92.5
w/o Local Weights	24.3	65.3	80.2	92.7
w/o Global Weight	24.7	65.6	80.0	92.6
Ours	**26.9**	**67.0**	**81.3**	**92.9**

TABLE VIII

THE ABLATION RESULTS (%) OF CROSS-DOMAIN ALIGNMENT TRAINING STRATEGY. COMPARED WITH TRAINING COSD OR SOCD SUB-TASK, TRAINING BOTH SUB-TASKS SIMULTANEOUSLY GETS HIGHER ACCURACY

COSD	SOCD	CroDom Dataset		
		$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$
✓	×	51.7	81.0	90.8
×	✓	60.6	81.7	90.1
✓	✓	**61.0**	**83.9**	**91.9**

2) Ablation on Probabilistic Correlation Module: We test different algorithm variants on the PF-PASCAL test set to evaluate the contribution of each proposed key component to semantic alignment accuracy. Tab. VII proves that the probabilistic correlation module helps to improve the PCK score and build more accurate alignment, as it incorporates multi-scale information to infer alignment. When we remove the confidence estimation block, the probabilistic correlation module would degenerate to a direct summation for multi-scale information like MMNet [17], rather than adaptive aggregation, which leads to the reduction of PCK. It is worth noting that both local weights and the global weight in the probabilistic correlation module are beneficial to semantic alignment accuracy. Employing them together has higher PCK than using only one of them.

3) Ablation on Training Strategy: We test various algorithm variants on the CroDom dataset to investigate the importance of COSD matching and SOCD matching in the training strategy of cross-domain semantic alignment. Tab. VIII shows that if training only one sub-task, PCK scores are less than that of training both sub-tasks at the same time, indicating that the combination of COSD matching and SOCD matching is beneficial to alignment accuracy. In addition, Fig. 11 presents warped target images based on estimated semantic alignments,

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on February 07,2024 at 02:17:05 UTC from IEEE Xplore. Restrictions apply.

Fig. 10. The visualized alignment errors of keypoints. The first column lists source images and columns 2-6 are target images, where the dots are ground-truth keypoints, while boxes are estimated keypoints by mapping source keypoints through the predicted semantic alignments. Green lines depict alignment errors.

Fig. 11. Visual comparison of the warping qualities of training COSD, SOCD, and both two sub-tasks. The last column shows the best warping quality, which qualitatively reflects the most accurate dense semantic alignments.

where warping quality qualitatively reflects dense alignment accuracy. And we can find that COSD matching and SOCD matching cannot reconstruct object contours well, and their warped images have some distortions and artifacts, while simultaneously training both sub-tasks can alleviate these problems and produce accurate and smooth warping results.

4) Improvements to SFNet: We train SFNet [24] with the landmark loss as a baseline. In order to verify the effectiveness of our proposed components, we add the enhancement transformer module to SFNet, and replace its original multi-scale multiplication of correlation maps with our probabilistic correlation module. Fig. 12 presents the PCK-score curves of the baseline and its variants. The results show that both the enhancement transformer module and the probabilistic correlation module can improve the PCK scores of the baseline. Using both modules together can bring more alignment accuracy to the baseline.

F. Analysis of Complexity and Runtime

Our enhancement module is inspired by ViT [30], which attends long-range information. We analyze the complexity

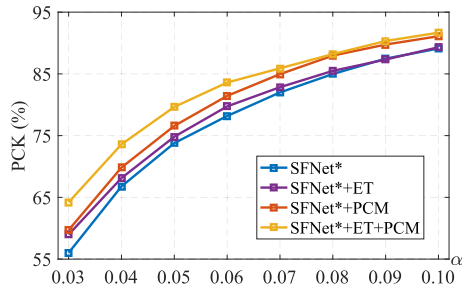


Fig. 12. Improving SFNet [24] using our enhancement transformer (ET) and probabilistic correlation module (PCM). $*$ means to retrain the network with the landmark loss. Both ET and PCM have a gain effect on SFNet, and using both has higher PCK scores, indicating better alignment accuracy.

TABLE IX
ANALYSIS OF ALGORITHMS WITH VARIOUS ENHANCEMENT MODULES. COLUMNS 2 AND 3 PRESENT PCK SCORES ON THE PF-PASCAL DATASET. COLUMNS 4-7 SHOW GPU-MEMORY COST, TIME-CONSUMING, THE AMOUNT OF FLOATING POINT ARITHMETICS, AND THE TOTAL NUMBER OF NETWORK PARAMETERS

Algorithm Variants	PF-PASCAL 0.05	PF-PASCAL 0.10	Mem.	Time	FLOPs	Params
ViT-based [30]	5.0	19.3	-	-	-	-
Pre-processing+ViT	80.7	92.8	8841MB	**0.189s**	1.257T	932.0M
Ours	**81.3**	**92.9**	**4673MB**	0.196s	**1.247T**	**363.7M**

and runtime of the algorithms with different enhancement modules, as shown in Tab. IX. The network complexity involves memory usage, the amount of floating point arithmetic's (FLOPs), and the number of network parameters. We can see from Tab. IX that the ViT-based method has low alignment accuracy with low PCK scores, because it ignores the importance of neighborhood during feature enhancement. Adding our pre-processing can address this problem, so we set the pre-processing and ViT-based method as the baseline. Different from the baseline, our method takes advantage of a sharing mechanism that shares part of network parameters, which can alleviate the network complexity. From Tab. IX, we can see that the baseline takes $8,841$MB memory and

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on February 07,2024 at 02:17:05 UTC from IEEE Xplore. Restrictions apply.

Fig. 13. Style transfer. B' is the average warping of B based on the semantic alignments between A and B. It guides to generate the stylized image C, achieving a style transfer from B to A.

has 932.0M network parameters. In contrast, our method only needs 4, 673MB memory and 363.7M parameters, meanwhile maintaining good alignment accuracy. For time consumption, our method takes 0.196s to process an image pair.

G. Limitations

Although our semantic alignment method achieves satisfactory performance on five cross-object datasets and a novel cross-domain dataset, it still has some limitations. We present several limitations and corresponding solutions as follows. First, our method obtains performance gain from the proposed network structure, but ignores the influence of training data, pre-trained features, and loss constraints. Building upon our method and exploring data augmentation, feature fine-tuning, and more complex loss might bring further improve alignment accuracy. Second, in the cross-object scenario, our method relies on landmark labels as supervision signals, similar to many other methods. However, this requires tedious manual labeling of landmarks, limiting a large number of potential images available on the Internet for training the network. The self-supervised training manner might address this limitation since it does not require hand-annotated labels.

H. Application Demonstrations

This section shows that our method can be effectively used for three applications: style transfer, semantic mask transfer, and correspondence-based image morphing. All results are better viewed in electronic form.

1) Style Transfer: An interesting application of our approach is to transfer a photo to a reference artistic style. Users can render their own photos with their favorite artwork (e.g., an oil painting or a watercolor) for sharing and entertainment. During style transfer, our method contributes to searching target locations. Specifically, to transfer the local style of image B to the area with the same semantics in image A, our method can look for the corresponding area based on semantic consistency. As shown in Fig. 13, we first warp B to an examplar B' using the estimated alignment field, which is semantically aligned with A while retaining the style of B, and then use it to guide generative adversarial networks (here we choose the network of Zhang et al. [19]) to transfer B's style to A. The final stylized images are presented in the right column of Fig. 13, which have B's style and A's structure.

Fig. 14. Foreground mask transfer. Given an image pair (A, B) and A's foreground mask, we predict B's foreground by warping the mask from A to B, according to the estimated semantic alignments between A and B.

Fig. 15. Alignment-based image morphing. The two original images are warped and fused into a morph sequence (20%,40%, 60%, 80%) based on our estimated semantic alignments, achieving a smooth morph that gradually transforms one image to another.

2) Semantic Mask Transfer: The foreground mask is a kind of semantic label, and its manual annotation is tedious. To overcome this problem, we can map a known foreground mask from one image to another semantically similar image, using the estimated dense semantic alignment field. As shown in Fig. 14, for the input image pair (A, B) that have challenging photometric and geometric variations but contain the same semantic content, our semantic alignment approach successfully transfers the foreground label. If a more accurate mask is required, users can further fine-tune the warped mask, significantly reducing labor costs when compared to labeling from scratch.

3) Alignment-Based Image Morphing: Alignment-based image morphing targets the production of a smooth and continuous morph animation that gradually transforms one image to another. Typically, motion paths are defined between corresponding points and then interpolated into dense smooth trajectories. Based on these trajectories, the input images are gradually warped and blended to produce an animation. The morph sequence is shown in the second row of Fig. 15. Since the quality of corresponding points is crucial for the success of the morph, manual alignment tagging is a reliable way to guarantee effective image morphing [5]. In contrast, we here utilize the semantic alignments obtained by our method to replace the user-provided alignments in the semi-automated

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on February 07,2024 at 02:17:05 UTC from IEEE Xplore. Restrictions apply.

image morphing method of Liao et al. [5], achieving automatic and smooth morphing.

V. Conclusion and Future Work

In this paper, we proposed a unified neural network architecture to address two types of semantic alignment issues, namely cross-object semantic alignment and cross-domain semantic alignment. Experimental results verified the effectiveness of our network architecture, in which the proposed enhancement transformer can provide features with global awareness, and a novel probabilistic correlation module can adaptively couple multi-scale information by confidence learning, thus improving semantic alignment accuracy. Our method achieves competitive performance on five standard cross-object semantic alignment benchmarks. Moreover, we introduced a new training strategy to learn cross-domain semantic alignment, which outperforms the SOTA method. Furthermore, we built the first cross-domain alignment dataset and explored the applications of semantic alignment. In the future, we would like to take advantage of diffusion models to expand the scale of the CroDom dataset, including more object categories, domains, and images. In addition, we would study the cross-modal semantic alignment between text and image modalities, which is more difficult than cross-domain alignment because of a huge modality gap.

References

[1] H. Zhang, X. Ye, S. Chen, Z. Wang, H. Li, and W. Ouyang, "The farther the better: Balanced stereo matching via depth-based sampling and adaptive feature refinement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4613–4625, Jul. 2022.

[2] H. Dai, X. Zhang, Y. Zhao, H. Sun, and N. Zheng, "Adaptive disparity candidates prediction network for efficient real-time stereo matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 3099–3110, May 2022.

[3] R. Zhao, R. Xiong, Z. Ding, X. Fan, J. Zhang, and T. Huang, "MRDFlow: Unsupervised optical flow estimation network with multi-scale recurrent decoder," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4639–4652, Jul. 2022.

[4] S. Liu, K. Luo, A. Luo, C. Wang, F. Meng, and B. Zeng, "ASFlow: Unsupervised optical flow learning with adaptive pyramid sampling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4282–4295, Jul. 2022.

[5] J. Liao, Y. Yao, L. Yuan, G. Hua, and S. B. Kang, "Visual attribute transfer through deep image analogy," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–15, Aug. 2017.

[6] J. Huang, J. Liao, and S. Kwong, "Semantic example guided image-to-image translation," *IEEE Trans. Multimedia*, vol. 23, pp. 1654–1665, 2021.

[7] Z. Zhang, Z. Wang, Z. Lin, and H. Qi, "Image super-resolution by neural texture transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7974–7983.

[8] M. He, D. Chen, J. Liao, P. V. Sander, and L. Yuan, "Deep exemplar-based colorization," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–16, Aug. 2018.

[9] I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional neural network architecture for geometric matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 39–48.

[10] P. H. Seo, J. Lee, D. Jung, B. Han, and M. Cho, "Attentive semantic alignment with offset-aware correlation kernels," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 349–364.

[11] S. Huang, Q. Wang, S. Zhang, S. Yan, and X. He, "Dynamic context correspondence network for semantic alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2010–2019.

[12] Y. Liu, L. Zhu, M. Yamada, and Y. Yang, "Semantic correspondence as an optimal transport problem," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4462–4471.

[13] J. Min, J. Lee, J. Ponce, and M. Cho, "Learning to compose hypercolumns for visual correspondence," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 346–363.

[14] S. Li, K. Han, T. W. Costain, H. Howard-Jenkins, and V. Prisacariu, "Correspondence networks with adaptive neighbourhood consensus," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10193–10202.

[15] J. Y. Lee, J. DeGol, V. Fragoso, and S. N. Sinha, "PatchMatch-based neighborhood consensus for semantic correspondence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13148–13158.

[16] X. Li, D. Fan, F. Yang, A. Luo, H. Cheng, and Z. Liu, "Probabilistic model distillation for semantic correspondence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7501–7510.

[17] D. Zhao, Z. Song, Z. Ji, G. Zhao, W. Ge, and Y. Yu, "Multi-scale matching networks for semantic correspondence," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3334–3344.

[18] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[19] P. Zhang, B. Zhang, D. Chen, L. Yuan, and F. Wen, "Cross-domain correspondence learning for exemplar-based image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5142–5152.

[20] J. L. Long, N. Zhang, and T. Darrell, "Do convnets learn correspondence?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1601–1609.

[21] T. Zhou, P. Krähenbühl, M. Aubry, Q. Huang, and A. A. Efros, "Learning dense correspondence via 3D-guided cycle consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 117–126.

[22] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, and J. Sivic, "Neighbourhood consensus networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 1658–1669.

[23] C. Barnes, E. Shechtman, D. B. Goldman, and A. Finkelstein, "The generalized patchmatch correspondence algorithm," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2010, pp. 29–43.

[24] J. Lee, D. Kim, J. Ponce, and B. Ham, "SFNet: Learning object-aware semantic correspondence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2273–2282.

[25] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2019, pp. 1–12.

[27] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," Tech. Rep., 2018.

[28] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 213–229.

[29] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6877–6886.

[30] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–13.

[31] W. Jiang, W. Zhou, and H. Hu, "Double-stream position learning transformer network for image captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7706–7718, Nov. 2022.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[33] B. Ham, M. Cho, C. Schmid, and J. Ponce, "Proposal flow: Semantic correspondences from object proposals," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 7, pp. 1711–1725, Jul. 2018.

[34] J. Min, J. Lee, J. Ponce, and M. Cho, "SPair-71k: A large-scale benchmark for semantic correspondence," 2019, *arXiv:1908.10543*.

[35] N. Kolkin, J. Salavon, and G. Shakhnarovich, "Style transfer by relaxed optimal transport and self-similarity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10043–10052.

[36] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 694–711.

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on February 07,2024 at 02:17:05 UTC from IEEE Xplore. Restrictions apply.

[37] Y.-C. Chen, P.-H. Huang, L.-Y. Yu, J.-B. Huang, M.-H. Yang, and Y.-Y. Lin, "Deep semantic matching with foreground detection and cycle-consistency," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 347–362.

[38] J. Min, J. Lee, J. Ponce, and M. Cho, "Hyperpixel flow: Semantic correspondence with multi-layer neural features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3394–3403.

[39] P. Truong, M. Danelljan, F. Yu, and L. Van Gool, "Probabilistic warp consistency for weakly-supervised semantic correspondences," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8698–8708.

[40] M. Aygün and O. Mac Aodha, "Demystifying unsupervised semantic correspondence estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 125–142.

[41] J. Min and M. Cho, "Convolutional Hough matching networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2939–2949.

[42] S. Cho, S. Hong, S. Jeon, Y. Lee, K. Sohn, and S. Kim, "CATs: Cost aggregation transformers for visual correspondence," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 9011–9023.

[43] J. Kim et al., "Semi-supervised learning of semantic correspondence with pseudo-labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19667–19677.

[44] S. Kim, J. Min, and M. Cho, "TransforMatcher: Match-to-match attention for semantic correspondence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8687–8697.

[45] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013.

[46] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker, "Universal correspondence network," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 2414–2422.

[47] B. Ham, M. Cho, C. Schmid, and J. Ponce, "Proposal flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3475–3484.

[48] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2006.

[49] T. Taniai, S. N. Sinha, and Y. Sato, "Joint recovery of dense correspondence and cosegmentation in two images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4246–4255.

[50] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing: Label transfer via dense scene alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1972–1979.

[51] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, pp. 303–308, Sep. 2009.

[52] M. J. Wilber, C. Fang, H. Jin, A. Hertzmann, J. Collomosse, and S. Belongie, "BAM! The behance artistic media dataset for recognition beyond photography," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1211–1220.

Huaiyuan Xu (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Tianjin University in 2014, 2017, and 2022, respectively.

He is currently a Post-Doctoral Research Fellow with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong, China. His current research interests include robotic perception, computer vision, and deep learning.

Jing Liao (Member, IEEE) received the B.Eng. degree from the Huazhong University of Science and Technology, and the dual Ph.D. degree from Zhejiang University and The Hong Kong University of Science and Technology.

She has been an Assistant Professor with the Department of Computer Science, City University of Hong Kong, since September 2018. Prior to that, she was a Researcher with the Visual Computing Group, Microsoft Research Asia, from 2015 to 2018. Her current research interests include the fields of computer graphics, computer vision, image/video processing, digital art, and computational photography.

Huaping Liu (Senior Member, IEEE) received the Ph.D. degree from Tsinghua University, in 2004.

He is currently a Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His current research interests include robot perception and learning. He was a recipient of the National Science Fund for Distinguished Young Scholars. He served as the Area Chair of Robotics Science and Systems for several times. He is a Senior Editor of *The International Journal of Robotics Research*.

Yuxiang Sun (Member, IEEE) received the bachelor's degree from the Hefei University of Technology, Hefei, China, in 2009, the master's degree from the University of Science and Technology of China, Hefei, in 2012, and the Ph.D. degree from The Chinese University of Hong Kong, Shatin, Hong Kong, in 2017.

He is currently a Research Assistant Professor with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong. His current research interests include autonomous driving, deep learning, robotics and autonomous systems, and semantic scene understanding.

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on February 07,2024 at 02:17:05 UTC from IEEE Xplore. Restrictions apply.