Multi-granularity Correlation Refinement for Semantic Correspondence

1st Zhen Liang✉
College of Computer and Information Science
Southwest University
Chongqing, P.R. China
zhenliang0628@gmail.com

2nd Enyu Che
College of Computer and Information Science
Southwest University
Chongqing, P.R. China
enyuche@gmail.com

3rd Guoqiang Xiao
College of Computer and Information Science
Southwest University
Chongqing, P.R. China
gqxiao@swu.edu.cn

4th Jingwei Qu*
College of Computer and Information Science
Southwest University
Chongqing, P.R. China
qujingwei@swu.edu.cn

***Abstract*—Semantic correspondence aims to establish dense correspondences between semantically similar images. Multi-level image features have been commonly used in recent studies due to their rich information. However, this approach poses a challenging problem of how to distinguish the importance of multiple similarity scores for each candidate match. Moreover, the introduction of the level dimension increases ambiguous matches. To address these challenges, we develop a Multi-granularity Inter-level Attention-based Matching (MIAM) network. Specifically, multi-scale inter-level self-attention, conditioned on correlation patches of various sizes, is proposed to adjust the effect of similarities from different levels on building correspondences. Next, we introduce a dual dimensional re-weighting strategy to further alleviate the ambiguity issue. Based on the convolutional aggregation of the multi-level scores along the spatial and level dimensions, this strategy strengthens positive matches while suppressing negative ones. In the thorough evaluation on three semantic correspondence benchmarks, MIAM achieves competitive performance compared to popular methods. Our algorithm is available at https://github.com/2000LZZ/MIAM.**

***Index Terms*—Semantic correspondence, multi-granularity, multi-level correlation**

I. Introduction

Semantic correspondence aims to establish pixel-wise, locally consistent correspondences between images that contain different instances of the same category. This task is crucial in numerous computer vision applications, such as 3D reconstruction [20]. It is more challenging than matching images just taken under different geometric settings due to the large intra-class appearance variations and geometric differences between the above semantically similar images [14], [22], [23].

Remarkable success of convolutional neural networks (CNNs) in various domains has led to the development of CNN-based methods for semantic correspondence [2], [26], [27]. These methods follow a pipeline similar to the classical matching, including feature extraction, cost aggregation, and flow estimation. Some methods focus on extracting multi-level image features [18], while others propose different representations of correspondences in the third stage [14], [24]. Given the importance of similarity scores (*i.e.*, correlation map) in matching tasks, recent approaches have shifted their attention to improving the quality of correlation map, *e.g.*, designing high-dimensional convolutional [13], [25] or Transformer-based [2], [11] aggregation. Nonetheless, the progress in addressing two critical challenges is still limited.

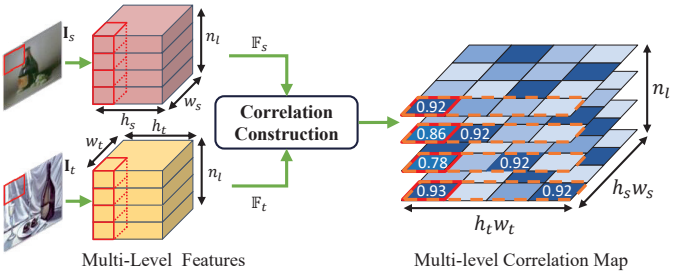


Fig. 1. A multi-level correlation map contains not only multiple similarity scores for each candidate match, but also more ambiguous matches (*e.g.*, matches with the same score but from different levels).

Previous studies have demonstrated that multi-level image features can aid in establishing dense correspondences. However, richer features also pose an inherent issue. In the resulting multi-level correlation map, the importance of each candidate match's multiple similarity scores varies (as illustrated in Fig. 1). It is crucial to adjust the effect of similarities from different levels based on their importance to ensure the quality of correspondences. Furthermore, the introduction of the level dimension raises more ambiguous matches, making it more difficult to distinguish between positive and negative matches in the search for the optimal correspondences.

To tackle the aforementioned challenges, we present a Multi-granularity Inter-level Attention-based Matching (MIAM) network for semantic correspondence (Fig. 2). Our

*Corresponding author.

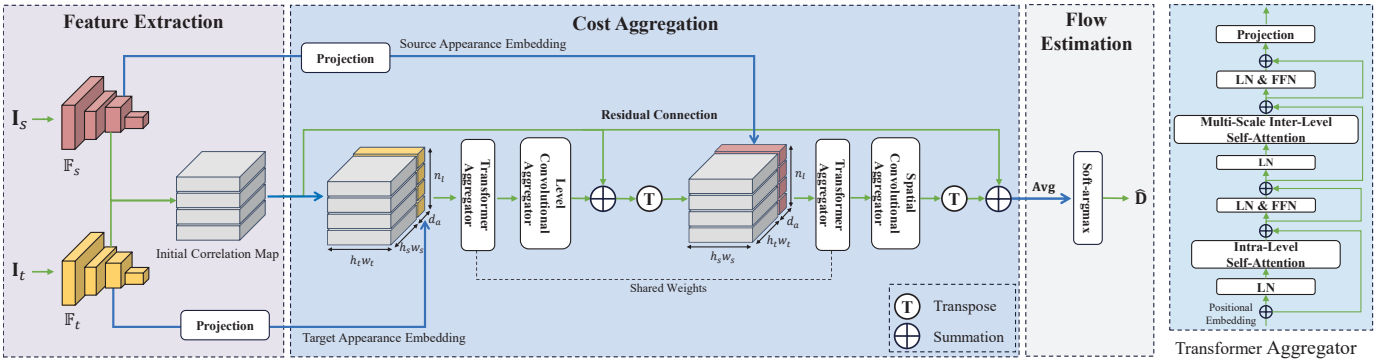


Fig. 2. Overall architecture of MIAM. First, the multi-level features of the given image pair $(\mathbf{I}_s, \mathbf{I}_t)$ are generated to compute the multi-level correlation map $\mathbf{C}$. Next, MIAM refines $\mathbf{C}$ by a sequential combination of the Transformer aggregator $\mathcal{T}$, the level convolutional aggregator $\mathcal{C}_l$, and the spatial convolutional aggregator $\mathcal{C}_s$. Finally, the refined correlation map $\mathbf{C}''$ is used to estimate dense correspondences between the images. The Transformer aggregator integrates the intra-level self-attention and the multi-scale inter-level self-attention with LN and FFN to refine correlation maps across spatial and level dimensions.

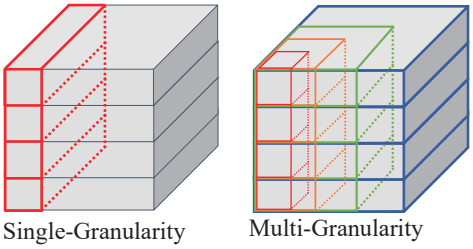


Fig. 3. Conceptual difference between CATs [2] and ours. Compared to a single patch size, multiple patch sizes enrich the granularity of the basic attention element.

solution incorporates a multi-scale inter-level self-attention mechanism, where we treat similarity scores in a correlation patch as an individual element for attention and consider different patch sizes to enrich the type of the attention element. The importance of similarities at different levels is captured from their interactions along the level dimension, and is then utilized to adaptively adjust their effects on building correspondences. To further address the issue of increasing ambiguities, we introduce a dual dimensional re-weighting strategy in MIAM. This strategy generates two attention maps by convolutional aggregation of the scores in the minimum and maximum patches along the spatial and level dimensions, respectively. The scores are then re-weighted using the attention maps, which strengthens positive matches while suppressing negative ones. Compared to the previous single patch size [2], [11], the multiple patch sizes form a multi-granularity refinement for the correlation map (Fig. 3). In experiments on three popular semantic correspondence benchmarks, MIAM achieves competitive performance in comparison with other methods, both quantitatively and qualitatively.

II. Methodology

The proposed Multi-granularity Inter-level Attention-based Matching (MIAM) network follows the classical three stages of semantic correspondence (Fig. 2): feature extraction, cost aggregation, and flow estimation.

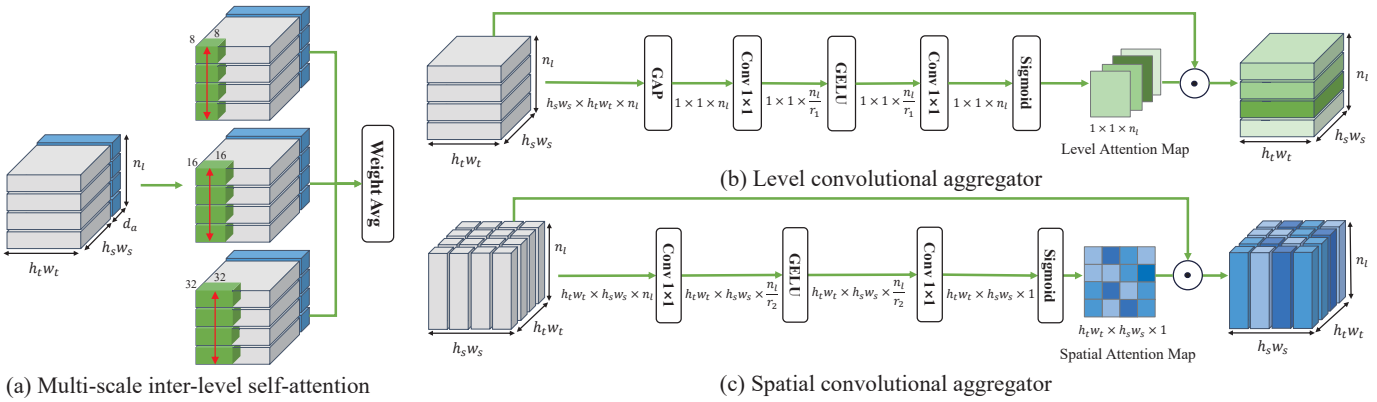
A. Feature Extraction

Given a pair of images to be established dense correspondences, referred to as the source and target images $\mathbf{I}_s, \mathbf{I}_t \in \mathbb{R}^{H\times W\times 3}$, a CNN ϕ is utilized to extract the respective multi-level features. Taking $\mathbf{I}_s$ as an example, a sequence of feature maps from different layers is extracted by the feature backbone ϕ. From these maps, several are selected and bi-linearly interpolated to the same spatial size $h \times w$. These resized feature maps create a multi-level feature set $\mathbb{F}_s = \{\mathbf{F}_s^k \in \mathbb{R}^{h_s\times w_s\times d_k}\}_{k=1}^{n_l}$, where $\mathbf{F}_s^k$ denotes a feature map at the k-th level, d_k is the number of channels of $\mathbf{F}_s^k$, n_l is the number of levels, and the subscript s is used only to distinguish the two images, in other words, $h_s = h_t = h$ and $w_s = w_t = w$.

B. Cost Aggregation over Multi-granularity Refinement

The multi-level correlation map $\mathbf{C} \in \mathbb{R}^{h_s w_s\times h_t w_t\times n_l}$ is then computed based on the two multi-level features $\mathbb{F}_s, \mathbb{F}_t$. Specifically, given the source and target feature maps $\mathbf{F}_s^k, \mathbf{F}_t^k$ at the k-th level, the k-th level of $\mathbf{C}$ is computed as $\mathbf{C}_{i,j,k} = \mathbf{F}^k_{s_{\mathbf{i},:}} \cdot \mathbf{F}^k_{t_{\mathbf{j},:}}$, where $\mathbf{i}, \mathbf{j} \in \mathbb{R}^2$, $i = \mathrm{id}(\mathbf{i})$, $j = \mathrm{id}(\mathbf{j})$, $\mathrm{id}(\cdot)$ is a bijection function that maps a 2D spatial position of a feature map of shape $h \times w$ to an integer index in $\{1, \ldots, hw\}$. In this way, each feature map is treated as $h * w$ local features, and $\mathbf{C}_{i,j,k}$ reveals the similarity score between the two local features $\mathbf{F}^k_{s_{\mathbf{i},:}}$ and $\mathbf{F}^k_{t_{\mathbf{j},:}}$. Next, the correlation map $\mathbf{C}$ is refined by the multi-granularity inter-level attention.

Multi-scale Inter-level Self-attention To explore the impact of similarities from different levels on establishing correspondences, we propose a multi-scale inter-level self-attention mechanism. As illustrated in Fig. 4(a), the correlation map $\mathbf{C}_{:,:,k} \in \mathbb{R}^{h_s w_s\times h_t w_t}$ at each level is transformed into a sequence of flattened patches with a fixed size, *i.e.*, $\mathbb{R}^{h_s w_s\times h_t w_t} \rightarrow \mathbb{R}^{n_p\times p^2}$, where (p, p) is the size of each patch, and $n_p = h_s w_s * h_t w_t / p^2$ is the resulting number of patches. Then n_l patches located at the same position across all levels form the input sequence $\mathbf{X} \in \mathbb{R}^{n_l\times p^2}$ for attention, *i.e.*, the similarity scores in each patch are treated as an attention element. Therefore, the multi-level correlation map



(a) Multi-scale inter-level self-attention

(b) Level convolutional aggregator

(c) Spatial convolutional aggregator

Fig. 4. Illustration of the multi-granularity refinement. (a) The multi-scale inter-level self-attention refines the correlation map by considering the similarity scores of each patch as an attention element and designing multiple patch sizes. Dual dimensional re-weighting strategy consists of (b) the level convolutional aggregator and (c) the spatial convolutional aggregator, which disambiguates similarity scores along the spatial and level dimensions.

$\mathbf{C}$ is transformed to n_p sequences, each with the length of n_l and the dimension of p^2, *i.e.*, $\mathbb{R}^{h_s w_s \times h_t w_t \times n_l} \rightarrow \mathbb{R}^{n_p \times n_l \times p^2}$. The multi-head self-attention (MHSA) of the n_p sequences is computed iteratively:

$$\begin{aligned} \mathbf{Q} = \mathbf{X}\mathbf{W}_Q,\ \mathbf{K} = \mathbf{X}\mathbf{W}_K,\ \mathbf{V} = \mathbf{X}\mathbf{W}_V \\ \mathrm{SA_h}(\mathbf{X}) = \mathrm{softmax}\Big(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_K}}\Big)\mathbf{V} \\ \mathrm{MHSA}(\mathbf{X}) = [\mathrm{SA_h}(\mathbf{X})]_{\mathrm{h}=1}^{n_h}\mathbf{W}_O \end{aligned} \tag{1}$$

where $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{p^2 \times d_K}$, $\mathbf{W}_V \in \mathbb{R}^{p^2 \times d_V}$, $\mathbf{W}_O \in \mathbb{R}^{n_h d_V \times p^2}$, $[\cdot]$ denotes the concatenation operation, h is the head index, and n_h is the number of heads. In this paper, we have omitted the details of bias terms for brevity.

Furthermore, multiple patch sizes are designed to form the multi-scale attention. The patch size determines the number of the similarity scores in a patch. Using various patch sizes can enrich the basic element type for attention. Thus, the correlation map $\mathbf{C}$ undergoes multi-scale attention with three different patch sizes $\mathbb{P} = \{8, 16, 32\}$, and the three outputs $\mathbb{C} = \{\tilde{\mathbf{C}}_i \in \mathbb{R}^{n_{p_i} \times n_l \times {p_i}^2}\}_{i=1}^{3}$ are reshaped back to $h_s w_s \times h_t w_t \times n_l$ to compute their weighted sum:

$$\hat{\mathbf{C}} = \sum_{i=1}^{3} \beta_i \tilde{\mathbf{C}}_i \tag{2}$$

where $\hat{\mathbf{C}} \in \mathbb{R}^{h_s w_s \times h_t w_t \times n_l}$ denotes the updated correlation map, $p_i \in \mathbb{P}$, and β_i is the learnable weight of each output. By mining various interactions between the similarities along the level dimension, the above self-attention can adaptively adjust the impact of similarities from different levels, thus reducing ambiguous matches.

We further design a Transformer-based aggregator $\mathcal{T}$ to refine the multi-level correlation map. To achieve this, we introduce a multi-head intra-level self-attention block [2] before the proposed inter-level self-attention block. The intra-level attention captures the interactions between similarity scores in a single-level correlation map. We then add a feed-forward network (FFN) after each MHSA block. The FFN block contains two linear layers with a GELU non-linearity. Besides, layer normalization (LN) is applied before the MHSA and FFN blocks for more effective optimization, a technique widely used in current Transformer implementations. Residual connections are applied after each block, and necessary reshaping operations are performed on the correlation map.

Dual Dimensional Re-weighting To further disambiguate the correlation map, a dual dimensional re-weighting strategy is introduced (Fig. 4). It complements the above attention mechanism by analyzing the scores in the minimum and maximum size patches, *i.e.*, $(1, 1)$ and (hw, hw).

Based on the concept that convolution operations combine cross-channel and spatial information [1], [28], we implement a level convolutional aggregator $\mathcal{C}_l$ and a spatial convolutional aggregator $\mathcal{C}_s$ to infer a level attention map $\mathbf{A}_l \in \mathbb{R}^{1 \times 1 \times n_l}$ and a spatial attention map $\mathbf{A}_s \in \mathbb{R}^{hw \times hw \times 1}$, respectively:

$$\begin{aligned} \mathbf{A}_l &= \sigma_2\Big(\mathbf{W}_2\sigma_1\big(\mathbf{W}_1\mathrm{GAP}(\mathbf{C})\big)\Big) \\ \mathbf{A}_s &= \sigma_2\big(\mathbf{W}_4\sigma_1(\mathbf{W}_3\mathbf{C})\big) \end{aligned} \tag{3}$$

where GAP denotes the global average pooling, σ_1 and σ_2 denote the GELU and sigmoid functions respectively, and $\mathbf{W}_1 \in \mathbb{R}^{n_l \times \frac{n_l}{r_1}}$, $\mathbf{W}_2 \in \mathbb{R}^{\frac{n_l}{r_1} \times n_l}$, $\mathbf{W}_3 \in \mathbb{R}^{n_l \times \frac{n_l}{r_2}}$ and $\mathbf{W}_4 \in \mathbb{R}^{\frac{n_l}{r_2} \times 1}$ indicate the weights of the convolution operation with a filter size of 1×1 for downscaling or upscaling the channel dimensions. The attention maps are then applied to re-weight the input correlation map:

$$\begin{aligned} \mathcal{C}_l(\mathbf{C}) &= \mathbf{A}_l \odot \mathbf{C} \\ \mathcal{C}_s(\mathbf{C}) &= \mathbf{A}_s \odot \mathbf{C} \end{aligned} \tag{4}$$

where $\odot$ denotes the element-wise multiplication. The level convolutional aggregator $\mathcal{C}_l$ uses the pooling operation to squeeze all scores in the largest patches of size (hw, hw) at each level. The spatial convolutional aggregator $\mathcal{C}_s$ fuses the multiple scores of each candidate match by focusing on the smallest patches of size $(1, 1)$. The resulting attention maps highlight where to focus along both level and spatial dimensions of the correlation map, thus enhancing positive matches while suppressing negative ones.

TABLE I

COMPARISON OF PCK@α_τ (%) ON PF-PASCAL, PF-WILLOW, AND SPAIR-71K. BOLD AND UNDERLINED NUMBERS INDICATE THE BEST AND THE SECOND BEST PERFORMANCE, RESPECTIVELY. "FEAT.-LEVEL": FEATURE-LEVEL, "FT. FEAT.": FINE-TUNE FEATURE. "(F)" AND "(T)" INDICATE FINE-TUNING-BASED TESTING AND TRANSFER TESTING, RESPECTIVELY.

Methods	Feat.-level	FT. feat.	Aggregation	PF-PASCAL PCK@α_{img} 0.05(F)	0.1(F)	0.15(F)	PF-WILLOW PCK@α_{obj} 0.1(T)	PCK@α_{kp} 0.1(T)	SPair-71k PCK@α_{obj} 0.1(F)	0.1(T)
CNNGeo [23]	Single	×	-	41.0	69.5	80.4	-	69.2	20.6	-
A2Net [27]	Single	×	-	42.8	70.8	83.3	-	68.8	22.3	-
WeakAlign [24]	Single	×	-	49.0	74.8	84.0	-	70.2	20.9	-
RTNs [10]	Single	×	-	55.2	75.9	85.2	-	71.9	25.7	-
NC-Net [26]	Single	✓	4D Convolution	54.3	78.9	86.0	-	67.0	20.1	26.4
DCC-Net [8]	Single	×	4D Convolution	55.6	82.3	90.5	-	73.8	-	26.7
ANC-Net [15]	Single	×	4D Convolution	-	86.1	-	-	-	-	<u>28.7</u>
CHM [17]	Single	✓	6D Convolution	<u>80.1</u>	91.6	94.9	79.4	69.6	46.3	**30.1**
SFNet [14]	Multi	×	-	53.6	81.9	90.6	-	74.0	28.2	-
HPF [18]	Multi	-	RHM	60.1	84.8	92.7	-	74.4	28.2	-
GSF [9]	Multi	×	2D Convolution	65.6	87.8	95.9	-	**78.7**	36.1	-
DHPF [19]	Multi	×	RHM	75.7	90.7	95.0	77.6	71.0	37.3	27.4
SCOT [16]	Multi	-	OT-RHM	63.1	85.4	92.7	-	<u>76.0</u>	35.6	-
MMNet [29]	Multi	✓	-	77.6	89.1	94.3	-	-	40.9	-
CATs [2]	Multi	✓	Transformer	75.4	<u>92.6</u>	<u>96.4</u>	79.2	69.0	49.9	27.1
VAT [7]	Multi	✓	4D Conv. & Trans.	78.2	92.3	96.2	**81.6**	-	**55.5**	-
TMatcher [11]	Multi	✓	Transformer	**80.8**	91.8	-	76.0	65.3	<u>53.7</u>	**30.1**
MIAM	Multi	✓	Transformer	77.5	**93.6**	**96.9**	<u>79.8</u>	70.0	50.5	28.2

Multi-granularity Refinement The above aggregators are arranged in series to refine the multi-level correlation map (Fig. 2). The process is designed based on two common augmentation approaches for the correlation map [2], [7]: (1) adding an appearance embedding from the image features to suppress noise; (2) swapping the order of the two input images to impose bidirectional matching consistency.

$$\begin{aligned}\mathbf{C}' &= \mathcal{C}_l\Big(\mathcal{T}_1\big([\mathbf{C}, \psi(\mathbb{F}_t)] + \mathbf{E}_1\big)\Big) + \mathbf{C} \\ \mathbf{C}'' &= \mathcal{C}_s\Big(\mathcal{T}_2\big([\mathbf{C}'^{\top}, \psi(\mathbb{F}_s)] + \mathbf{E}_2\big)\Big)^{\top} + \mathbf{C}\end{aligned} \tag{5}$$

Two Transformer-based aggregators, $\mathcal{T}_1$ and $\mathcal{T}_2$, with shared parameters are utilized, and combined with the level convolutional aggregator $\mathcal{C}_l$ and the spatial convolutional aggregator $\mathcal{C}_s$, respectively. For $(\mathcal{T}_1, \mathcal{C}_l)$, the appearance embeddings $\psi(\mathbb{F}_t)$ from target image features are concatenated with the initial correlation map $\mathbf{C}$, and a learnable positional embedding $\mathbf{E}_1 \in \mathbb{R}^{(h_s w_s + d_a) \times h_t w_t \times n_l}$ is added as input, where $\psi : \mathbb{R}^{h \times w \times d_i} \rightarrow \mathbb{R}^{hw \times d_a}$ denotes the linear projection. The first two dimensions of the output $\mathbf{C}' \in \mathbb{R}^{h_s w_s \times h_t w_t \times n_l}$ are then swapped to reflect the target-source order[1]. The source appearance embeddings $\psi(\mathbb{F}_s)$ are concatenated with the output and then added with another positional embedding $\mathbf{E}_2 \in \mathbb{R}^{(h_t w_t + d_a) \times h_s w_s \times n_l}$ as input for $(\mathcal{T}_2, \mathcal{C}_s)$. The resulting output $\mathbf{C}'' \in \mathbb{R}^{h_s w_s \times h_t w_t \times n_l}$ is the final refined result. Residual connections are applied to both outputs.

[1]The superscript $\top$ in Eq. (5) denotes the swapping operation on the first two dimensions of the correlation map.

C. Flow Estimation & Training Objective

After refining the multi-level correlation map, we average it along the level dimension. Then, we transform the averaged correlation map into a dense flow field $\hat{\mathbf{D}}$ from the source to target images by the kernel soft-argmax operation [14]. Finally, we introduce a loss based on the Euclidean distance between the predicted and ground-truth flow fields to train our model:

$$\mathcal{L} = ||\mathbf{D} - \hat{\mathbf{D}}|| \tag{6}$$

The flow field $\mathbf{D}$ is constructed using ground-truth keypoints following the protocol in [18]. It is assumed that the ground-truth keypoints are provided for image pairs.

III. EXPERIMENTS

The performance of our MIAM is evaluated by comparing it with several state-of-the-art methods. Additionally, ablation studies are conducted on crucial components of MIAM to analyze their effectiveness.

A. Experimental Settings & Implementation Details

The experiments are conducted on three popular benchmarks: PF-PASCAL [5], PF-WILLOW [4] and SPair-71k [18]. We follow the general evaluation protocol to ensure a fair comparison. For PF-PASCAL and PF-WILLOW, we train MIAM on the training split of PF-PASCAL and evaluate on the test splits of both. For SPair-71k, MIAM is trained on the training split and then evaluated on the test. The quality of semantic correspondence is evaluated using the standard evaluation metric, probability of correct keypoints (PCK), which depends on the threshold $\alpha_\tau \cdot \max(h_\tau, w_\tau)$.

TABLE II
MEMORY AND RUN-TIME COMPARISON.

Methods	Patch size	Mem. (GB)	Time (ms)	PF-P	PF-W
VAT [7]	$(1, 1)$	2.2	59.6	92.3	81.6
CATs [2]	$(384, 1)$	0.9	10.9	92.6	79.2
MIAM	$(32, 32)$	1.0	11.6	93.1	79.3
	$(16, 16)$	0.9	12.9	93.4	79.4
	$(8, 8)$	0.9	19.5	93.3	79.4
	$\{8, 16, 32\}$	1.0	22.5	93.6	79.8

$\tau \in \{\text{img}, \text{obj}, \text{kp}\}$ indicates that the threshold relies on the height and width of image, object bounding box, or bounding box of keypoints, respectively. $\alpha_\tau \in \{0.05, 0.1, 0.15\}$ is a tolerance factor.

ResNet-101 [6] pre-trained on ImageNet [3] is adopted as the feature backbone ϕ. The spatial sizes of the input images and the multi-level features are set to $H = W = 256$ and $h = w = 16$, respectively. The number of feature levels is set to $n_l = 8$. We set the number of layers in the Transformer aggregator $\mathcal{T}$ to 1, the dimensions of the QKV vectors and the number of heads in the MHSA blocks to $d_K = d_V = 48$ and $n_h = 6$, and the dimensions of the appearance embeddings to $d_a = 128$. The ratio parameters of the two convolutional aggregators, $\mathcal{C}_s$ and $\mathcal{C}_l$, are both set to $r_1 = r_2 = 8$. MIAM is implemented by PyTorch [21], and is optimized via AdamW [12] with an initial learning rate of 3×10^{-5}, which gradually decreases during training. These hyperparameters are fixed for all experiments. All experiments are run on a single NVIDIA GeForce RTX 4090 GPU and an Intel Core i9-13900K CPU.

B. Evaluation Results

Quantitative Results The quantitative results are illustrated in Tab. I. Overall, MIAM achieves promising performance on all three benchmarks. Methods that use multi-level features perform better due to richer semantic information. MIAM outperforms these methods with the highest PCK of 93.6% ($\alpha_{\text{img}} = 0.1$) on PF-PASCAL. In comparison to VAT [7], MIAM achieves a lower PCK for SPair-71k. This could be attributed to VAT's use of high-dimensional convolution and Swin Transformer to extend local receptive fields during the cost aggregation stage. However, this design results in longer run-time and larger memory requirements, which we will discuss in the following efficiency experiments. We also assess the correspondence quality of the model trained on PF-PASCAL directly on SPair-71k (last column of Tab. I). When combined with the results on PF-WILLOW, MIAM demonstrates competitive transferability.

Efficiency Evaluation Further experiments are conducted to evaluate the efficiency of MIAM by profiling the required memory and run-time (Tab. II). We report the GPU memory consumed by a pair of images and the inference time for their correspondences. The multi-scale inter-level self-attention captures richer interactions between similarity scores, but it also requires more memory and longer run-time due to the combination of multiple patch sizes. Using a single size results in a greater increase in running efficiency, but sacrifices some performance.

Fig. 5. Qualitative results on PF-PASCAL. Red lines indicate failed cases.

TABLE III
ABLATION STUDIES OF MIAM ON PF-PASCAL.

Methods	PCK@0.1
(a) Baseline	92.6
(b) Baseline+DDR	93.2
(c) Baseline+DDR+MIS	93.6

Qualitative Results Fig. 5 visualizes the predicted correspondences of several challenging image pairs. Compared to VAT [7] and CATs [2], MIAM is able to distinguish ambiguities and produce desired correspondences in cases involving repetitive patterns and background clutter.

C. Ablation Study

Key Components To evaluate the individual effect of MAIM's key components, we perform ablation studies on the **m**ulti-scale **i**nter-level **s**elf-attention (MIS) and the **d**ual **d**imensional **r**e-weighting (DDR) strategy. We define a baseline model by excluding these two components. The baseline model uses a single patch size $(hw + d_a, 1)$ to replace the multiple patch sizes $\mathbb{P}$. The experiments are conducted by successively incorporating each key component into the baseline model. The results presented in Tab. III demonstrate that the model's performance improves from **(a)** to **(b)** and from **(b)** to **(c)** as each component is added, highlighting the effectiveness of the two key components in refining the multi-level correlation map.

TABLE IV
ABLATION STUDIES ON CONVOLUTIONAL AGGREGATORS.

Order	PCK@0.1
Spatial-Level	93.4
Level-Spatial	93.6

Convolutional Aggregators To analyze the impact of the order of the two convolutional aggregators on the correspondence quality, we compare two ordering ways: spatial-level and level-spatial. The results in Tab. IV show that the level-first order (93.6%) performs slightly better than the spatial-first order (93.4%) with a difference of 0.2% on PF-PASCAL. The reason for this may be that the spatial convolutional aggregator implicitly embeds interactions between levels by fusing multiple scores of all candidate matches. The level-spatial order forms an explicit-implicit refining on the level dimension of the multi-level correlation map.

IV. CONCLUSION

This paper introduces a Multi-granularity Inter-level Attention-based Matching (MIAM) network for predicting image correspondences. MIAM utilizes a multi-scale inter-level self-attention, which is conditioned on multiple correlation patch sizes, to adaptively adjust the effect of similarities from different levels. Furthermore, we propose a dual dimensional re-weighting strategy for disambiguating multi-level similarities along the level and spatial dimensions. Experimental results demonstrate the effectiveness of MIAM.

ACKNOWLEDGMENT

This work was supported in part by Natural Science Foundation of Chongqing, China (No. CSTB2023NSCQ-MSX0881) and Fundamental Research Funds for the Central Universities (No. SWU-KR22032).

REFERENCES

[1] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yang, and F. Yu, "Dual aggregation transformer for image super-resolution," in *Proceedings of the International Conference on Computer Vision*, 2023, pp. 12 312–12 321.

[2] S. Cho, S. Hong, S. Jeon, Y. Lee, K. Sohn, and S. Kim, "Cats: Cost aggregation transformers for visual correspondence," *Proceedings of the Conference on Neural Information Processing Systems*, vol. 34, pp. 9011–9023, 2021.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2009, pp. 248–255.

[4] B. Ham, M. Cho, C. Schmid, and J. Ponce, "Proposal flow," in *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2016, pp. 3475–3484.

[5] ——, "Proposal flow: Semantic correspondences from object proposals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 7, pp. 1711–1725, 2017.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2016, pp. 770–778.

[7] S. Hong, S. Cho, J. Nam, S. Lin, and S. Kim, "Cost aggregation with 4d convolutional swin transformer for few-shot segmentation," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 108–126.

[8] S. Huang, Q. Wang, S. Zhang, S. Yan, and X. He, "Dynamic context correspondence network for semantic alignment," in *Proceedings of the International Conference on Computer Vision*, 2019, pp. 2010–2019.

[9] S. Jeon, D. Min, S. Kim, J. Choe, and K. Sohn, "Guided semantic flow," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 631–648.

[10] S. Kim, S. Lin, S. R. Jeon, D. Min, and K. Sohn, "Recurrent transformer networks for semantic correspondence," *Proceedings of the Conference on Neural Information Processing Systems*, vol. 31, 2018.

[11] S. Kim, J. Min, and M. Cho, "Transformatcher: Match-to-match attention for semantic correspondence," in *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2022, pp. 8697–8707.

[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[13] J. Y. Lee, J. DeGol, V. Fragoso, and S. N. Sinha, "Patchmatch-based neighborhood consensus for semantic correspondence," in *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2021, pp. 13 153–13 163.

[14] J. Lee, D. Kim, J. Ponce, and B. Ham, "Sfnet: Learning object-aware semantic correspondence," in *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2019, pp. 2278–2287.

[15] S. Li, K. Han, T. W. Costain, H. Howard-Jenkins, and V. Prisacariu, "Correspondence networks with adaptive neighbourhood consensus," in *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2020, pp. 10 196–10 205.

[16] Y. Liu, L. Zhu, M. Yamada, and Y. Yang, "Semantic correspondence as an optimal transport problem," in *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2020, pp. 4463–4472.

[17] J. Min and M. Cho, "Convolutional hough matching networks," in *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2021, pp. 2940–2950.

[18] J. Min, J. Lee, J. Ponce, and M. Cho, "Hyperpixel flow: Semantic correspondence with multi-layer neural features," in *Proceedings of the International Conference on Computer Vision*, 2019, pp. 3395–3404.

[19] ——, "Learning to compose hypercolumns for visual correspondence," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 346–363.

[20] A. Mustafa and A. Hilton, "Semantically coherent co-segmentation and reconstruction of dynamic scenes," in *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2017, pp. 422–431.

[21] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Proceedings of the Conference on Neural Information Processing Systems*, vol. 32, 2019.

[22] J. Qu, H. Ling, C. Zhang, X. Lyu, and Z. Tang, "Adaptive edge attention for graph matching with outliers," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2021, pp. 966–972.

[23] I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional neural network architecture for geometric matching," in *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2017, pp. 6148–6157.

[24] I. Rocco, R. Arandjelović, and J. Sivic, "End-to-end weakly-supervised semantic alignment," in *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2018, pp. 6917–6925.

[25] ——, "Efficient neighbourhood consensus networks via submanifold sparse convolutions," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 605–621.

[26] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, and J. Sivic, "Neighbourhood consensus networks," *Proceedings of the Conference on Neural Information Processing Systems*, vol. 31, 2018.

[27] P. H. Seo, J. Lee, D. Jung, B. Han, and M. Cho, "Attentive semantic alignment with offset-aware correlation kernels," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 349–364.

[28] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–19.

[29] D. Zhao, Z. Song, Z. Ji, G. Zhao, W. Ge, and Y. Yu, "Multi-scale matching networks for semantic correspondence," in *Proceedings of the International Conference on Computer Vision*, 2021, pp. 3354–3364.