GLOBAL-LOCAL BAYESIAN TRANSFORMER FOR SEMANTIC CORRESPONDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Cost aggregation is the key to finding semantic correspondence between a pair of similar images. Transformer-based cost aggregators have recently shown strong performance in obtaining high-quality correlation maps due to their capability of capturing long-range dependencies between matching points. However, such models are data-hungry and prone to over-fitting when training data is not sufficiently large. Besides, they easily incur incorrect matches when finding correspondences in the local semantic context. To address these issues, we propose a Global-Local Bayesian Transformer (GLBT) for cost aggregation. Specifically, GLBT introduces one global Bayesian self-attention module, whose weights are sampled from a learnable Bayesian posterior distribution, to mitigate over-fitting while modeling the long-range interaction from correlation maps. Furthermore, to model the short-range interaction between candidate matches, GLBT introduces another local Bayesian self-attention module, which factorizes both correlation maps and Bayesian attention weights into pairs of patches and conducts a matrix multiplication on individuals rather than a direct dot-product. Two self-attention modules are joined together to model the long-range and short-range interactions from correlation maps. Ultimately, GLBT is hierarchically aggregated for the refinement of correlation maps before feeding it to the flow estimator. We conduct extensive experiments to show the superiority of our proposed network to the state-of-the-art methods on datasets, including SPair-71k, PF-PASCAL, and PF-WILLOW.

1 INTRODUCTION

Establishing dense semantic correspondences between images is a fundamental problem facilitating many vision tasks, including semantic segmentation (Min et al., 2021; Xie et al., 2021), 3D reconstruction (Kokkinos & Kokkinos, 2021a;b; Li et al., 2020b), and optical flow estimation(Yang & Ramanan, 2019). In contrast to the classical pixel-wise correspondence problems (Kim et al., 2003) that require images to be geometrically normalized and aligned, semantic correspondence considers unconstrained image pairs, posing additional challenges from large intra-class variations in appearance and geometry.

Recent methods (Bristow et al., 2015; Cho et al., 2021; Zhao et al., 2021) for semantic correspondence generally follow the classical matching pipeline, including feature extraction, cost aggregation, and flow estimation. Some works (Rublee et al., 2011; Tola et al., 2010) attempted to find the semantic similarity between images by focusing on the feature extraction stage. These methods disregard the pixel-wise relationship between correlation features, resulting in sub-optimal performance. To overcome this issue, several methods (Jeon et al., 2020; Rocco et al., 2017; Truong et al., 2020b; Hong & Kim, 2021) introduced a regression network at the flow estimation stage to infer dense correspondences from correlation maps. However, such approaches rely on high-quality initial matching scores. Thereby, the latest methods (Min & Cho, 2021; Min et al., 2019a; Li et al., 2020a; Rocco et al., 2020; Min et al., 2020; Rocco et al., 2018b) have focused on designing an efficient cost aggregation module to improve the quality of correlation maps before feeding them into the flow estimation, proving the importance of cost aggregation networks.

The core of the cost aggregation stage is to produce reliable correlation maps via the refinement of matching scores. Some models (Min & Cho, 2021; Rocco et al., 2018b) refined the local consistent matches from the initial correlation maps with high-dimensional 4D or 6D convolutions. However,

such models lack the ability to achieve long-range context aggregation due to the inherently limited receptive fields. To tackle this problem, CATs (Cho et al., 2021) leveraged the vision transformer for cost aggregation to effectively refine the ambiguous matching scores in consideration of the global consensus. Nonetheless, it overlooks the spatial structure of the correlation map, leading to sub-optimal results. To further boost the performance, VAT (Hong et al., 2022) proposed a 4D Convolutional Swin Transformer as a cost aggregator to preserve the spatial structure of correlation maps, while providing an efficient self-attention to model long-range interaction between candidate matches. However, the existing Transformer-based cost aggregators (Hong et al., 2022; Casey et al., 2021; Cho et al., 2021) are infeasible to model the short-range pixel-to-pixel interaction, resulting in redundant noisy matches when dealing with the local semantic matches. In addition, since transformer architecture is prone to over-fitting, these transformer-based aggregators are data-hungry (Hassani et al., 2021), i.e., requiring enormous amounts of training data to obtain a good performance.

To address these limitations, we propose a Global-Local Bayesian Transformer (GLBT) cost aggregator for semantic correspondence. Inspired by BayesNN (Blundell et al., 2015), which applied a variational inference on the weights of a neural network to prevent over-fitting, our proposed GLBT introduces the Global-Local Bayesian Self-Attention (GLB-SA) into the transformer aggregator for capturing the long-range and short-range match-to-match interaction from correlation maps simultaneously. Compared to the raw self-attention in the transformer (Cho et al., 2021; Vaswani et al., 2017), which suffers from a data-hungry issue due to the operation of dense matrix-vector multiplication, GLBT leverages the sparse matrix factorization (Dao et al., 2019) on the self-attention operation to avoid over-fitting via a reduction in its learnable parameters. The proposed GLBT module is then leveraged to hierarchically aggregate the multi-level matching correspondences on the different semantic contexts, achieving the refinement of correlation maps. Consequently, the refined correlation maps are applied in the decoder to infer the semantic correspondences from image pairs.

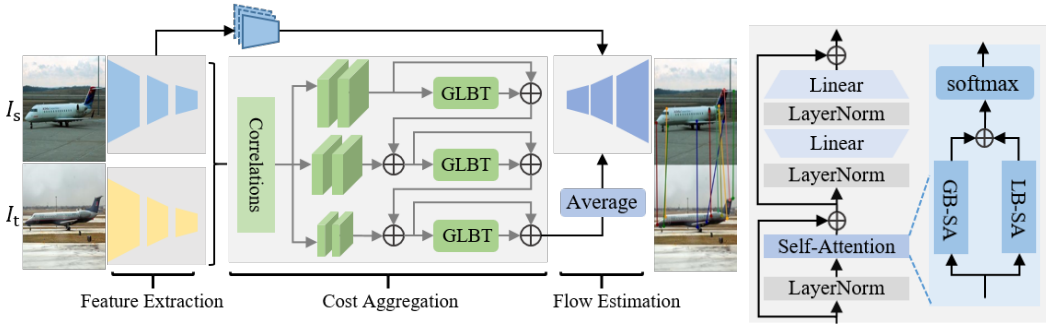
We validate the effectiveness of our GLBT method on public benchmark datasets (Ham et al., 2016; 2017; Min et al., 2019b). Extensive experimental results demonstrate that our proposed method for semantic correspondence outperforms the previous state-of-the-art methods on several benchmarks. We also provide a detailed ablation analysis to verify the main components in GLBT.

2 Related Work

Semantic Correspondence. Finding semantic correspondences between image pairs poses additional challenges to intra-class appearance and shape variations among different instances from the same object or scene category. To address these challenges, approaches to semantic correspondence can be roughly categorized into hand-crafted feature-based methods (Bay et al., 2006; Dalal & Triggs, 2005; Ham et al., 2016; Liu et al., 2011; LoweDavid, 2004; Rublee et al., 2011; Tola et al., 2010) and learnable feature-based methods (Choy et al., 2016; Kim et al., 2018; 2017; Lee et al., 2019; Li et al., 2020a; Rocco et al., 2017; Seo et al., 2018). Hand-crafted techniques leverage the low-level feature descriptors, such as SIFT (LoweDavid, 2004), HOG (Taniai et al., 2016), and DAISY (Tola et al., 2010), to measure dense correspondences, lacking the capture of high-level semantics.

To tackle this problem, most learnable techniques focus on building dense correspondences on high-level semantic features of deep convolutional neural networks, such as NC-Net (Rocco et al., 2018b), ANC-Net (Li et al., 2020a), and GOCor (Truong et al., 2020a). However, solely relying on the deep learnable features limits the performance of semantic correspondences due to the direct output of the similarity scores from the correlation maps. To address this issue, (Rocco et al., 2017) proposed a regression network to estimate the parameters from the matching features, coping with incorrect matches from the initial learnable features at the flow estimation stage. Their success encourages many variant methods, e.g., GSF (Jeon et al., 2020) and GLU-Net (Truong et al., 2020b), to directly regress semantic correspondences from the feature matches.

Cost Aggregation. To alleviate the requirement of high-quality initial matching scores, HPF (Min et al., 2020) introduced the RHM (Min et al., 2019a) cost aggregator into the learnable feature methods for geometric consistency enhancement. Later, numerous CNN-based feature-learnable variants (Min & Cho, 2021; Rocco et al., 2018b) utilized 4D or 6D convolution-based geometric matching algorithms to refine the local consistency of the initial correlation maps. Nonetheless, CNN-based aggregation networks fail to model global matches due to the limited receptive fields of convolutions. Transformer-based aggregators (Cho et al., 2021; Sun et al., 2021), which leveraged



(a) Overall Pipeline of Semantic Correspondence (b) GLBT

Figure 1: **Overall architecture of our proposed GLBT.** (a) The whole pipeline includes feature extraction, cost aggregation, and flow estimation. Given a pair of images, a ResNet101 (He et al., 2016) is used to extract multi-level features independently. At the cost aggregation stage, we construct the multi-level correlation maps via a cosine similarity and propose the GLBT to refine the correlation maps hierarchically. At the last flow estimation stage, we concatenate the resulting matching scores with the source features to obtain the final correspondence field. (b) The GLBT joints GB-SA and LB-SA to model the long-range and short-range interactions from rich semantic correlation maps.

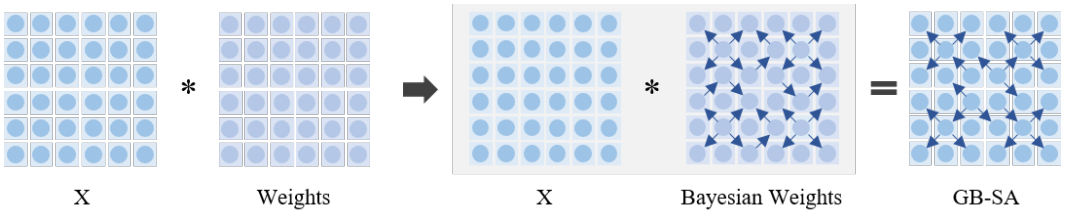
the self-attention mechanism (Vaswani et al., 2017) to capture the global match-to-match interaction from the initial correlation map, can solve this problem. However, such a self-attention is prone to introduce redundant noisy matches when modeling the short-range interaction in a small region, because it does not consider local contexts. Besides, Transformer-based deep networks starve huge amounts of training data to avoid over-confident decisions.

Bayesian Neural Network. Applying Bayesian approaches (Shridhar et al., 2019; Fan et al., 2021; Zhang et al., 2021) to neural networks is an alternative to mitigating the over-fitting issue by offering uncertainty estimates so that Bayesian Neural Networks (BayesNNs) can easily learn from small datasets and are robust to over-fitting. In the past years, several methods, such as Variational Inference (Blundell et al., 2015; Graves, 2011), Laplace Approximation (MacKay, 1992), and MC Dropout (Gal & Ghahramani, 2015; 2016), have been widely applied to estimate the parameter uncertainty, which is propagated for predictions. Instead of selecting a single point estimate, BayesNNs use the Bayes rule to average results over parameter values and thus have a strong reasoning ability.

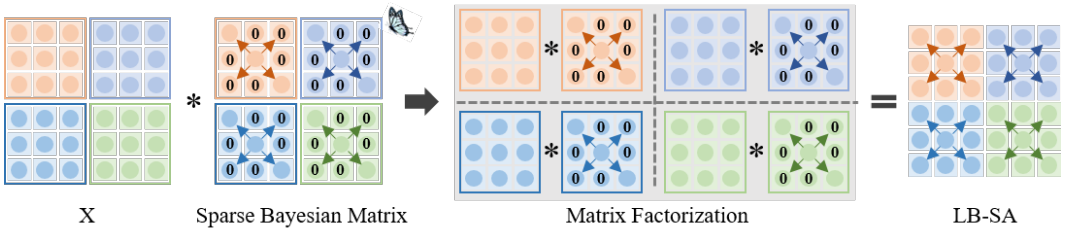
3 PRELIMINARY

Let $I^s \in \mathbb{R}^{H_s \times W_s \times 3}$ and $I^t \in \mathbb{R}^{H_t \times W_t \times 3}$ denote a pair of source and target images, respectively. The goal of dense semantic correspondence is to find the optimal f^* that generates a correspondence flow containing the offsets between corresponding keypoints in the two images, i.e., $\mathcal{K}^{pred} = f^*(I^s, I^t)$, where the correspondence flow $\mathcal{K}^{pred} = \{(\Delta x_i^s, \Delta y_i^s)\}_{i=1}^{H_s \times W_s}$ contains the predicted offsets for all pixels in the source image. Following previous works, we consider learning of f^* in the supervised setting. More specifically, we are given a dataset $\mathcal{D} = \{(I_j^s, I_j^t, \mathcal{K}_j^{gt})\}_{j=1}^{M}$ containing M image pairs and the associated ground-truth correspondence flows. Due to sparse annotations, the ground-truth flow $\mathcal{K}_j^{gt} = \{(\Delta x_i^s, \Delta y_i^s)\}_{i=1}^{H_s \times W_s}$ is only non-zero at a subset of locations. We aim to learn an approximate f_n by minimizing the distance between the predicted and the ground-truth correspondence flows: $f_M = \arg\min_f \frac{1}{M} \sum_{j=1}^{M} ||\Phi(f(I_j^s, I_j^t)) - \mathcal{K}_j^{gt}||$, where Φ is a logical metric that sets the offsets at locations without ground-truth offsets to zero.

The pipeline to design the function f involves several basic steps, including feature extraction, cost aggregation, and flow estimation. Specifically, dense feature maps $D^s \in \mathbb{R}^{H_s \times W_s \times C}$ and $D^t \in \mathbb{R}^{H_t \times W_t \times C}$ are extracted from each image pair I^s and I^t, respectively. Directly matching similarity between D^s and D^t without introducing any prior often undergoes ambiguous matches due to limited local repetitive patterns. To address this issue, cost aggregation techniques are employed to refine matches from initial correlation maps. The correspondence flow is, consequently, inferred from refined matching scores. Our approach follows this common framework for semantic correspondence.



(a) Global Bayesian Self-Attention (GB-SA)



(b) Local Bayesian Self-Attention (LB-SA)

Figure 2: **Global-Local Bayesian Self-Attention** is consist of (a) *GB-SA* and (b) *LB-SA*, modeling the long-range and short-range interactions between candidate matches respectively. Here, the left arrow visualizes the details of the computation process.

As shown in Figure 1(a), we follow the previous works (Min et al., 2021; 2020; Truong et al., 2020b) to construct the correlation maps, using cosine similarity, $\mathcal{C}(D^s_{\mathbf{x}}, D^t_{\mathbf{y}}) \in \mathbb{R}^{H_s \times W_s \times H_t \times W_t} = \frac{D^s_{\mathbf{x},:} \cdot D^t_{\mathbf{y},:}}{||D^s_{\mathbf{x},:}|| \cdot ||D^t_{\mathbf{y},:}||}$. The result is a 4D tensor, representing the initial matching scores between an image pair. To capture rich semantic information, 4D convolutions (Min & Cho, 2021; Rocco et al., 2018b) are employed to extract multi-scale features at different levels of a backbone network. However, such dense feature points are weak to identify global semantics alignment due to the limited receptive fields of convolutions. To address this issue, a cost aggregator is introduced to refine the correlation maps before feeding them for flow estimation. The current cost aggregators (Cho et al., 2021; Hong et al., 2022) leverage the transformer to refine correlation maps due to its global receptive fields. However, such methods discard the ability to capture the short-range interaction between candidate matches, leading to extra noisy matches when matching the semantic correspondences in a small region. Besides, it is data-hungry and requires large amounts of training data to avoid over-fitting.

4 Global-Local Bayesian Transformer

Given the limitations of existing methods, we propose a Global-Local Bayesian Transformer (GLBT) to refine the correlation maps by considering the local and global interactions between candidate matches simultaneously. As visualized in Figure 1(b), GLBT stacks a group of the Global-Local Bayesian Self-Attention (GLB-SA) module, layer normalization, and multilayer linear perceptron, to refine the final correlation matches:

$$\mathcal{C}' = \mathit{GLBT}(X), \tag{1}$$

where $\mathrm{X} \in \mathbb{R}^{L \times C}$ is the correlation map unfolded from the result of $\mathit{Conv4D}(\mathcal{C}(D^s, D^t))$, $L = H_s \times W_s \times H_t \times W_t$, and C denotes the channels.

Self-attention, which obtains key, value and query from the initial correlation map X, is the core of GLBT. Instead of using the standard self-attention mechanism (Vaswani et al., 2017), we introduce one Global Bayesian Self-Attention (GB-SA) in Figure 2(a) to model the global match-to-match interaction on the large semantic displacement, and another Local Bayesian Self-Attention (LB-SA) in Figure 2(b) to model the local match-to-match interaction on the small semantic displacement. Both are then joined together to reason about the final correlation maps at the same time.

4.1 Global Bayesian Self-Attention

The classical self-attention (Dosovitskiy et al., 2021) performs a dot-product on all pixels, prone to the issue of data-hungry (Liu et al., 2021; Yuan et al., 2021). BayesCNNs (Shridhar et al., 2019)

averages models sampled from the posterior distribution of convolution kernels and have the potential to prevent the requisite of large data and be robust to over-fitting. Inspired by this, we introduce a Global Bayesian Self-Attention (GB-SA), which directly operates the matrix-multiplication on the input and the Bayesian weight to learn the global interaction between candidate matches from the correlation maps.

Let θ denote the network parameters in the computation of the correlation map $X = h_\theta(I^s, I^t)$, and $W \in \mathbb{R}^d$ denote the parameters in a Bayesian self-attention module. Our Bayesian model considers W as a random variable and our goal is to infer the posterior distribution $p(W|\mathcal{D})$ and learn the parameters θ simultaneously. The whole proposed network can be viewed as the following probabilistic model:

$$p_\theta(\mathcal{K}^{gt}|I^s, I^t, W) = \mathcal{N}(\mathcal{K}^{gt}|\Phi(\mathcal{G}(h_\theta(I^s, I^t), W)), \sigma_0^2), \tag{2}$$

where $\mathcal{G}$ stands for the probability function in the GB-SA module, h_θ is the network computing the correlation map, and σ_0 is the standard deviation of the Gaussian distribution. To avoid yielding a slow convergence and prevent a strange local minima (Blundell et al., 2015), we use the mixture Gaussian distribution with zero mean for the prior distribution $p(W)$:

$$p(W) = \prod_{i=1}^{d} \mathcal{N}(W_i|0, \sigma_1^2) + (1-\pi)\mathcal{N}(W_i|0, \sigma_2^2), \tag{3}$$

where σ_1 and σ_2 correspond to the standard deviations of two Gaussian distributions, respectively.

To infer the Bayesian posterior distribution $p(W|\mathcal{D})$ on the weights in self-attention, we follow the variational inference procedure (Shridhar et al., 2019) to estimate an approximate variational posterior $q_\phi(W)$ by minimizing the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951):

$$\begin{aligned} \hat{\theta}, \hat{\phi} &= \arg\min_{\theta,\phi} \mathrm{KL}[q_\phi(W)||p_\theta(W|\mathcal{D})] \\ &= \arg\min_{\theta,\phi} \mathrm{KL}\left[q_\phi(W)\|p(W)\right] - \mathbb{E}_{q_\phi(W)}[\log p_\theta(\mathcal{D}|W)], \end{aligned} \tag{4}$$

where the likelihood $p_\theta(\mathcal{D}|W) = \prod_{j=1}^{M} p_\theta(\mathcal{K}_j^{gt}|I_j^s, I_j^t, W)$. In addition, we use a Gaussian distribution for the variational posterior, so the parameter $\phi = (\boldsymbol{\mu}, \boldsymbol{\sigma})$, where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\sigma}$ is the standard deviation vector.

To achieve the GB-SA operation visible in Figure 2(a), we sample the attention weights W_g^{bayes} from the learnable Bayesian posterior distribution $q_\phi(W|\mathcal{D})$ and, directly compute it and the correlation map X via the general matrix multiplication:

$$W_{GB} = \mathcal{G}(X, W_g^{bayes}) = X * W_g^{bayes}. \tag{5}$$

Recall the self-attention mechanism (Vaswani et al., 2017), the attention weight is a non-zero matrix with each row summing to one. Therefore, we leverage a softmax function to obtain attention weight $W_g = \mathrm{softmax}(W_{GB})$, and the resulting weight W_g is then used to refine the initial matching features. Such a GB-SA, which is leveraged as a regularisation on the weights of the network, can learn the global matches on small data and is robust to over-fitting.

4.2 Local Bayesian Self-Attention

Besides, the traditional self-attention (Dosovitskiy et al., 2021) is prone to introducing extra noisy matches when capturing the short-range interaction from the correlation maps in a small region. Inspired by the sliding windows used in convolutions, we introduce another Local Bayesian Self-Attention (LB-SA), which conducts the dot-product of the input and the sparse Bayesian weight according to the matrix factorization, to reason about the short-range matches from semantic context.

To achieve the LB-SA, we leverage the butterfly matrix (Dao et al., 2019) to generate a boolean matrix B and, sample the attention weight W_l^{bayes} from the learnable Bayesian posterior distribution $q_\phi(W|\mathcal{D})$ which is inferred by the similar rules as Equation 4. As shown in Figure 2(b), we employ the boolean matrix B to sparsify the Bayesian attention weight W_l^{bayes}:

$$A = B \odot W_l^{bayes}, \tag{6}$$

where $\odot$ is the Hadamard product of B and W_l^{bayes}, and the resulting A is a sparse Bayesian weight.

To capture the local correspondences on the limited receptive field, we leverage the matrix factorization technique to divide both the input X and the sparse Bayesian matrix A into n pairs of patches X_{ij} and A_{ij} with a window size $S \times S$, where $1 \leq i \leq n, 1 \leq j \leq n$ and $n = \frac{H_s}{S} = \frac{W_s}{S} = \frac{H_t}{S} = \frac{W_t}{S}$ Afterwards, each pair of sub-matrices is computed separately via the matrix multiplication, to generate the final Bayesian attention weight W_{LB}. Let $\mathcal{L}$ denote the function in the LB-SA module, we have:

$$\begin{aligned} W_{LB} &= \mathcal{L}(X, A) = X * A \\ &= \left[\begin{array}{c|c} X_{11} & X_{12} \\ \hline X_{21} & X_{22} \end{array}\right] * \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right] = \left[\begin{array}{c|c} X_{11} * A_{11} & X_{12} * A_{12} \\ \hline X_{21} * A_{21} & X_{22} * A_{22} \end{array}\right]. \end{aligned} \tag{7}$$

Compared to direct matrix-multiplication, such a process has a strong capability of modeling the local patterns while reducing the computational complexity. To efficiently model the long-range and short-range interactions between candidate matches from correlation maps, the resulting local attention weight W_{LB} is integrated with the global attention weight W_{GB} to obtain the final global-local attention weight $W_{gl} = \mathrm{softmax}(W_{GB} + W_{LB})$ in our proposed GLBT. Consequently, the GLBT is hierarchically aggregated as a cost aggregator to refine the initial correlation maps before feeding it into the decoder for flow estimation.

5 Experiments

5.1 Experimental Settings and Implementation Details

Datasets. We conduct comprehensive experiments on three widely-used benchmark datasets for semantic correspondence, including SPair-71k (Min et al., 2019b), PF-PASCAL (Ham et al., 2017) and PF-WILLOW (Ham et al., 2016). The SPair-71k dataset contains 70,958 image pairs with diverse variations in viewpoint and scale, splitting into 53,340 pairs for training, 5,384 pairs for validation and 12,234 pairs for testing. The PF-PASCAL dataset contains 1,351 image pairs from 20 categories, augmented to 2,940 training pairs, 308 validation pairs and 299 testing pairs. The PF-WILLOW dataset contains 900 image pairs from 4 categories, used for testing.

Evaluation Metric. The percentage of correct keypoints (PCK) is the standard evaluation metric for category-level matching. Given a pair of predicted keypoint $\mathcal{K}^{\mathrm{pred}}$ and ground-truth keypoint $\mathcal{K}^{\mathrm{gt}}$, PCK computes the ratio of correctly predicted keypoints by $PCK = \frac{1}{N}\sum_{i=1}^{N}[||\mathcal{K}_i^{\mathrm{pred}} - \mathcal{K}_i^{\mathrm{gt}}|| \leq \alpha \cdot \max(H, W)]$, where H and W denote height and width of an entire image or an object bounding box, and α is a threshold to tolerate the distance between the predicted keypoint and the ground-truth.

Implementation Details. We follow the recent method (Min et al., 2019a) to extract the features from the best sub-layers of ResNet101 (He et al., 2016) pre-trained on the ImageNet (Deng et al., 2009) dataset. In training process, batch-size is set to 8 for all experiments and AdamW (Kingma & Ba, 2015) with a weight decay of 0.05 is adopted for optimization. The data augmentation techniques introduced in (Cho et al., 2021) are also used in our method. The learning rate for backbone features is set to 1e-6. The learning rate for the cost aggregation layers is initialized as 1e-5 and gradually decreased during training. We train the model for 300 epochs. All experiments are implemented with PyTorch (Paszke et al., 2019) and our method costs 38.6 ms inference time on V100 GPUs.

5.2 Benchmark Results and Analysis

To provide a fair comparison of our proposed GLBT and other state-of-the-arts, including CNNGeo (Rocco et al., 2017), A2Net (Seo et al., 2018), NC-Net (Rocco et al., 2018b), WeakAlign (Rocco et al., 2018a), HPF (Min et al., 2019a), SCOT (Liu et al., 2020), DHPF (Min et al., 2020), CHM (Min & Cho, 2021), CATs (Cho et al., 2021), MMNet (Zhao et al., 2021), and VAT (Hong et al., 2022), we use the same backbone ResNet101 (He et al., 2016) to extract the features from a pair of images. All results are measured under the same PCK evaluation indications on the benchmark datasets.

Table 1 and Table 2 report the quantitative comparison of the proposed GLBT with the previous state-of-the-art methods on SPair-71k (Min et al., 2019b), PF-PASCAL (Ham et al., 2017), and PF-WILLOW (Ham et al., 2016) respectively. In Table 1, we find that the transformer-based cost aggregators outperform others by a wide margin, due to the capability of long-range matches for

Table 1: **Quantitative comparisons of different state-of-the-art methods evaluated on standard benchmarks (Ham et al., 2016; 2017; Min et al., 2019b). U** denotes unsupervised learning methods, **W** refers to weakly-supervised methods, and **F** represents fully-supervised methods. The backbone used here is ResNet101 (He et al., 2016). Higher PCK [%] is better. The best results are in bold, and the second-best results are underlined.

	Methods	Aggregation	SPair-71k PCK @ α_{bbox}	PF-PASCAL PCK @ α_{img}			PF-WILLOW PCK @ α_{bbox}		
			0.1	0.05	0.1	0.15	0.05	0.1	0.15
U	CNNgeo (Rocco et al., 2017)	2D Conv.	20.6	41.0	69.5	80.4	36.9	69.2	77.8
	A2Net (Seo et al., 2018)	2D Conv.	22.3	42.8	70.8	83.3	36.3	68.8	84.4
W	NC-Net (Rocco et al., 2018b)	4D Conv.	20.1	54.3	78.9	86.0	33.8	67.0	83.7
	WeakAlign (Rocco et al., 2018a)	2D Conv.	20.9	49.0	74.8	84.0	37.0	70.2	79.9
	RTNs (Kim et al., 2018)	2D Conv.	25.7	55.2	75.9	85.2	41.3	71.9	86.2
	DCC-Net (Huang et al., 2019)	4D Conv.	-	55.6	82.3	90.5	43.6	73.8	86.5
	PWarpC (Truong et al., 2022)	PWarpC	33.5	65.7	87.6	93.1	47.5	78.3	89.0
F	SFNet (Lee et al., 2019)	2D Conv.	24.0	59.0	84.0	92.0	46.3	74.0	84.2
	HPF (Min et al., 2019a)	RHM	28.2	60.1	84.8	92.7	45.9	74.4	85.6
	GSF (Jeon et al., 2020)	2D Conv.	36.1	65.6	87.8	95.9	49.1	78.7	90.2
	ANC-Net (Li et al., 2020a)	4D Conv.	-	-	86.1	-	-	-	-
	DHPF (Min et al., 2020)	RHM	37.3	75.7	90.7	95.0	49.5	77.6	89.1
	SCOT (Liu et al., 2020)	OT-RHM	35.6	63.1	85.4	92.7	47.8	76.0	87.1
	CHM (Min & Cho, 2021)	6D Conv.	46.3	-	91.6	94.9	52.7	79.4	87.5
	CATs (Cho et al., 2021)	Transformer	49.9	75.4	<u>92.6</u>	<u>96.4</u>	50.3	79.2	90.3
	MMNet (Zhao et al., 2021)	4D Conv.	50.4	77.6	91.6	95.9	-	-	-
	VAT (Hong et al., 2022)	Transformer	<u>55.5</u>	78.2	92.3	96.2	<u>52.8</u>	<u>81.6</u>	<u>91.4</u>
	GLBT	Transformer	**57.5**	**78.6**	**93.3**	**96.6**	**53.4**	**82.7**	**92.2**

Table 2: **Comparisons with the state-of-the-art methods on the SPair-71k (Min et al., 2019b) dataset.** All results are evaluated using PCK @ $\alpha_{\text{img}} = 0.1$. The best results are reported in bold.

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	dog	horse	mbike	person	plant	sheep	train	tv	all
CNNGeo (Rocco et al., 2017)	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 (Seo et al., 2018)	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
NC-Net (Rocco et al., 2018b)	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
WeakAlign (Rocco et al., 2018a)	22.2	17.6	41.9	15.1	38.1	27.4	27.2	31.8	12.8	26.8	22.6	14.2	20.0	22.2	17.9	10.4	32.2	35.1	20.9
HPF (Min et al., 2019a)	25.2	18.9	52.1	15.7	38.0	22.8	19.1	52.9	17.9	33.0	32.8	20.6	24.4	27.9	21.1	15.9	31.5	35.6	28.2
SCOT (Liu et al., 2020)	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 (Min et al., 2020)	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
CHM (Min & Cho, 2021)	49.6	29.3	68.7	29.7	45.3	48.4	39.5	64.9	20.3	60.5	56.1	46.0	33.8	44.3	38.9	31.4	72.2	55.5	46.3
CATs (Cho et al., 2021)	52.0	34.7	72.2	34.3	49.9	57.5	43.6	66.5	24.4	63.2	56.5	52.0	42.6	41.7	43.0	33.6	72.6	58.0	49.9
MMNet (Zhao et al., 2021)	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
VAT (Hong et al., 2022)	58.8	40.0	**75.3**	40.1	**52.1**	59.7	44.2	69.1	23.3	**75.1**	61.9	57.1	46.4	49.1	**51.8**	41.8	80.9	70.1	55.5
GLBT	**61.4**	**43.6**	75.0	**43.7**	49.9	**66.9**	**54.6**	**69.9**	26.1	72.8	**62.7**	**57.9**	**47.9**	**53.9**	49.8	**41.4**	**82.2**	**74.7**	**57.5**

self-attention in the transformer. Compared to the previous best transformer-based VAT, the overall performance of our GLBT surpasses it by 2.0% @ $\alpha_{\text{bbox}} = 0.1$, 1.0% @ $\alpha_{\text{img}} = 0.1$ and 1.1% @ $\alpha_{\text{img}} = 0.1$, on SPair-71k, PF-PASCAL, and PF-WILLOW, respectively. Moreover, we also compare the results of each class on SPair-71k in Table 2. GLBT achieves the best performance in most categories, such as aeroplane, bike and boat, because it integrates both global and local self-attention to learn the long-range and short-range matches between images when refining the matching scores.

Figure 3 provides the visual comparison of results obtained from GLBT and the recent state-of-the-art methods, namely VAT (Hong et al., 2022), MMNet (Zhao et al., 2021) and CATs (Cho et al., 2021). The visual examples demonstrate that GLBT can match more accurate points between a pair of images than other methods. The results also present that GLBT has smaller offsets than others for the correspondences between image pairs, further validating the effectiveness of our proposed method.

5.3 Ablation Study and Analysis

In this section, we provide an ablation analysis to investigate the importance of the cost aggregation stage during the entire pipeline. We also show the details of our proposed GLBT, including one Global Bayesian Self-Attention (GB-SA) and another Local Bayesian Self-Attention (LB-SA). For a fair comparison, we conduct all ablation study experiments with the same backbone ResNet101 (He et al., 2016) and each experiment is trained from scratch under the same settings.

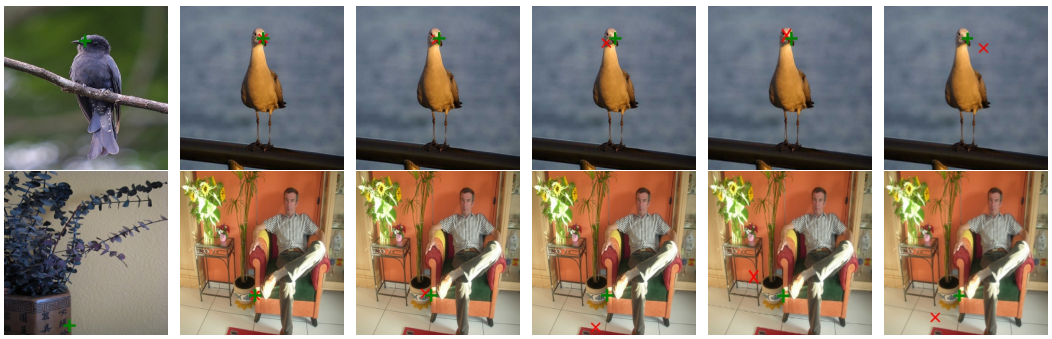


Figure 3: **Qualitative comparison of the recent state-of-the-art methods evaluated on SPair-71k (Min et al., 2019b),** including VAT (Hong et al., 2022), MMNet (Zhao et al., 2021) and CATs (Cho et al., 2021). All results are generated from the same model which is evaluated using PCK @ $\alpha_{\text{img}} = 0.1$. For each image pair, "+" is the groundtruth point and "$\times$" is the predicted key point. The closer distance between two signs corresponds to the better results.

Table 3: **Ablation results on the overall architecture.** "Feat", "Aggr" and "Flow" denote feature extraction, cost aggregation and flow estimation, respectively. SPair-71k and PF-WILLOW employs PCK @ α_{bbox} for evaluation, while PF-PASCAL adopts PCK @ α_{img}.

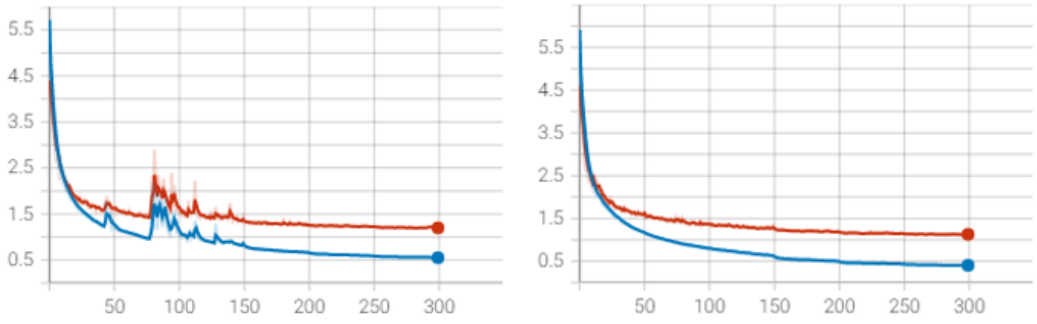
Feat	Aggr	Flow	SPair-71k	PF-PASCAL			PF-WILLOW		
			$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.15$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.15$
✓	✗	✗	44.5	69.7	86.9	94.2	39.1	73.6	86.7
✓	✗	✓	46.7	70.3	89.2	95.0	40.6	75.0	88.0
✓	✓	✗	54.6	76.5	92.7	96.2	52.9	80.8	90.8
✓	✓	✓	57.5	78.6	93.3	96.6	53.4	82.7	92.2

Overall Pipeline. Table 3 explores the impact of three modules, including feature extraction, cost aggregation, and flow estimation, for semantic correspondence. To validate that cost aggregation plays an essential role in the whole pipeline, we conduct ablation studies based on the different combinations of these modules. The results shown in Table 3 report the performance of involved models on SPair-71k (Min et al., 2019b), PF-PASCAL (Ham et al., 2017), and PF-WILLOW (Ham et al., 2016) in terms of PCK evaluation indicators with different thresholds. The results summarize that cost aggregation network contributes the most improvements to the final performance.

Effect on GLB Self-Attention. As visible in Table 4, we explore the effectiveness of the Global and Local Bayesian Self-Attention (GLB-SA) for transformer-based cost aggregation network, on the SPair-71k (Min et al., 2019b), PF-PASCAL (Ham et al., 2017), and PF-WILLOW (Ham et al., 2016) benchmark datasets in terms of PCK @ $\alpha = 0.1$. The baseline method adopts the Global Self-Attention (G-SA) Vaswani et al. (2017) based transformer to model the long-range matches between images for the refinement of correlation maps at the cost aggregation stage. To process the local semantic matches, we leverage matrix factorization (Ocker & Buice, 2021; Shah et al., 2015) to implement the Local Self-Attention (L-SA). Table 4 shows that the result of L-SA outperforms the G-SA. Besides, the Global-Local Self-Attention (GL-SA), which is a combination of G-SA and L-SA, has a better performance than both G-SA and L-SA. To further investigate the effects of

Table 4: **Comparison of various self-attention mechanisms for Transformer-based Cost Aggregator.** G-SA and L-SA mean the standard global and local self-attentions respectively. GL-SA represents the global-local self-attention. GB-SA and LB-SA denote the global and local Bayesian self-attentions respectively. GLB-SA is the global-local Bayesian self-attention.

	SPair-71k $\alpha_{\text{bbox}} = 0.1$	PF-PASCAL $\alpha_{\text{img}} = 0.1$	PF-WILLOW $\alpha_{\text{bbox}} = 0.1$
G-SA	54.6	90.6	80.0
GB-SA	55.3	91.8	81.1
L-SA	55.0	91.1	80.6
LB-SA	56.4	92.9	82.0
GL-SA	55.6	92.1	81.4
GLB-SA	57.5	93.3	82.7



(a) GLT (b) GLBT

Figure 4: **Convergence analysis of GLT vs. GLBT for cost aggregator.** (a) *GLT* shows the loss curve obtained by the global and local self-attention based transformer for cost aggregator, and (b) *GLBT* indicates the loss curve which is trained using the global and local Bayesian self-attention for cost aggregator. The blue and red curves represent the training and testing datasets, respectively.

Bayesian self-attention for the transformer cost aggregator, we conduct the extra ablation experiments, including G-SA *vs.* the Global Bayesian Self-Attention (GB-SA), L-SA *vs.* the Local Bayesian Self-Attention (LB-SA), and the GL-SA *vs.* GLB-SA, respectively. Compared to the results shown in Table 4, we find that the application of the Bayesian inference to self-attention in the transformer outperforms the non-Bayesian self-attention, because such a Bayesian self-attention mechanism acts like a regularization. Among them, our proposed GLB-SA achieves the best performance on the refinement of the correlation, further validating its effect on finding semantic correspondence.

Effect on Over-fitting for GLBT. To verify that the proposed Bayesian self-attention for the GLBT model can alleviate over-fitting, Figure 4 compares the loss curves obtained by the GLT and the GLBT. As shown in Figure 4(a), the loss curve of GLT fluctuates up and down in 50-150 epochs. The fluctuations are caused by the large intra-class variations in appearance and geometry for unconstrained image pairs. Figure 4(b) reports the loss curve of the GLBT, which is much more smooth than the GLT. We find that such a Bayesian self-attention can be regarded as a regularization mechanism to prevent the transformer-based model from over-fitting, when refining the correlation maps of challenging image pairs.

Memory and Run-time. Table 5 compares the memory and run-time of DHPF (Min et al., 2020), CHM (Min & Cho, 2021), CATs (Cho et al., 2021), MMNet (Zhao et al., 2021), VAT (Hong et al., 2022) and GLBT. For a fair comparison, all methods employ the backbone ResNet101 for feature extraction, and the results are obtained using the same machine. Compared to other methods, GLBT and VAT methods leverage transformer-based cost aggregators, exploit larger resolution and more memory than others, and surpass other methods by a large margin. We also find that compared to the previous state-of-the-art method VAT, our proposed method outperforms it in terms of PCK @ $\alpha = 0.1$, while reducing the memory and run-time by 0.4 GB and 18.7 ms, respectively.

Table 5: **Memory and run-time comparison.**

Methods	Resolution	Memory (GB)	Run-time[ms]
DHPF	240×240	1.6	57.7
CHM	240×240	1.6	47.2
CATs	256×256	1.9	34.5
MMNet	224×320	1.2	86.0
VAT	512×512	3.8	57.3
GLBT	512×512	3.4	38.6

6 Conclusion

In this paper, we have proposed a global-local Bayesian Transformer-based cost aggregation network, dubbed GLBT, for semantic correspondence. It integrates the global and local Bayesian self-attentions to infer the long-and-short range relationship between the correlation matches based on Bayes' rule, achieving both global and local match-to-match interaction at the same time. We have demonstrated that our proposed method outperforms the existing state-of-the-art by a large margin on public benchmark datasets. Moreover, we have also conducted extensive ablation studies to validate the effect of our proposed global-local Bayesian self-attention which is applied for Transformer-based cost aggregator. We hope that our findings can inspire further research work for other domains.

REFERENCES

Herbert Bay, Tinne Tuytelaars, and Luc Gool. Surf: Speeded up robust features. In *ECCV*, 2006.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *ICML*, 2015.

Hilton Bristow, Jack Valmadre, and Simon Lucey. Dense semantic correspondence where every pixel is a classifier. In *ICCV*, 2015.

Evan Casey, Víctor Pérez, and Zhuoru Li. The animation transformer: Visual correspondence via segment matching. In *ICCV*, 2021.

Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Cats: Cost aggregation transformers for visual correspondence. In *NIPS*, 2021.

Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. In *NIPS*, 2016.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

Tri Dao, Albert Gu, Matthew Eichhorn, Atri Rudra, and Christopher Ré. Learning fast algorithms for linear transforms using butterfly factorizations. In *ICML*, 2019.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Xinjie Fan, Shujian Zhang, Bo Chen, and Mingyuan Zhou. Bayesian attention modules. In *NIPS*, 2021.

Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. In *NIPS*, 2015.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

Alex Graves. Practical variational inference for neural networks. In *NIPS*, 2011.

Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow. In *CVPR*, 2016.

Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. *TPAMI*, 2017.

Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi. Escaping the big data paradigm with compact transformers. *arXiv*, 2021.

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

Sunghwan Hong and Seungryong Kim. Deep matching prior: Test-time optimization for dense correspondence. In *ICCV*, October 2021.

Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. *ECCV*, 2022.

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

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

Junhwan Kim, Kolmogorov, and Zabih. Visual correspondence using energy minimization and mutual information. In *ICCV*, 2003.

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

Seungryong Kim, Stephen Lin, SANG RYUL JEON, Dongbo Min, and Kwanghoon Sohn. Recurrent transformer networks for semantic correspondence. In *NIPS*, 2018.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Filippos Kokkinos and Iasonas Kokkinos. Learning monocular 3d reconstruction of articulated categories from motion. In *CVPR*, 2021a.

Filippos Kokkinos and Iasonas Kokkinos. To the point: Correspondence-driven monocular 3d category reconstruction. In *NIPS*, 2021b.

Solomon Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 1951.

Junghyup Lee, Dohyung Kim, Jean Ponce, and Bumsub Ham. Sfnet: Learning object-aware semantic correspondence. In *CVPR*, 2019.

Shuda Li, Kai Han, Theo W. Costain, Henry Howard-Jenkins, and Victor Prisacariu. Correspondence networks with adaptive neighbourhood consensus. In *CVPR*, 2020a.

Xueting Li, Sifei Liu, Shalini De Mello, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Online adaptation for consistent mesh reconstruction in the wild. In *NIPS*, 2020b.

Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *TPAMI*, 2011.

Yahui Liu, Enver Sangineto, Wei Bi, Nicu Sebe, Bruno Lepri, and Marco Nadai. Efficient training of visual transformers with small datasets. In *NIPS*, 2021.

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

G LoweDavid. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

David J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 1992.

Juhong Min and Minsu Cho. Convolutional hough matching networks. In *CVPR*, 2021.

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

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

Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Learning to compose hypercolumns for visual correspondence. In *ECCV*, 2020.

Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *ICCV*, 2021.

Gabriel Ocker and Michael Buice. Tensor decompositions of higher-order correlations by nonlinear hebbian plasticity. In *NIPS*, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NIPS*, 2019.

I. Rocco, R. Arandjelović, and J. Sivic. Convolutional neural network architecture for geometric matching. In *CVPR*, 2017.

I. Rocco, R. Arandjelović, and J. Sivic. End-to-end weakly-supervised semantic alignment. In *ECCV*, 2018a.

I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, and J. Sivic. Neighbourhood consensus networks. In *NIPS*, 2018b.

I. Rocco, R. Arandjelović, and J. Sivic. Efficient neighbourhood consensus networks via submanifold sparse convolutions. In *ECCV*, 2020.

Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, 2011.

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

Parikshit Shah, Nikhil Rao, and Gongguo Tang. Sparse and low-rank tensor decomposition. In *NIPS*, 2015.

Kumar Shridhar, Felix Laumann, and Marcus Liwicki. A comprehensive guide to bayesian convolutional neural network with variational inference. *arXiv*, 2019.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *CVPR*, 2021.

Tatsunori Taniai, Sudipta N. Sinha, and Yoichi Sato. Joint recovery of dense correspondence and cosegmentation in two images. In *CVPR*, 2016.

Engin Tola, Vincent Lepetit, and Pascal Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *TPAMI*, 2010.

Prune Truong, Martin Danelljan, Luc V Gool, and Radu Timofte. Gocor: Bringing globally optimized correspondence volumes into your neural network. In *NIPS*, 2020a.

Prune Truong, Martin Danelljan, and Radu Timofte. Glu-net: Global-local universal network for dense flow and correspondences. In *CVPR*, 2020b.

Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Probabilistic warp consistency for weakly-supervised semantic correspondences. In *CVPR*, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

Guo-Sen Xie, Huan Xiong, Jie Liu, Yazhou Yao, and Ling Shao. Few-shot semantic segmentation with cyclic memory network. In *ICCV*, 2021.

Gengshan Yang and Deva Ramanan. Volumetric correspondence networks for optical flow. In *NIPS*, 2019.

Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. In *ICCV*, 2021.

Shujian Zhang, Xinjie Fan, Bo Chen, and Mingyuan Zhou. Bayesian attention belief networks. In *ICML*, 2021.

Dongyang Zhao, Ziyang Song, Zhenghao Ji, Gangming Zhao, Weifeng Ge, and Yizhou Yu. Multi-scale matching networks for semantic correspondence. In *ICCV*, 2021.