Efficient Semantic Matching with Hypercolumn Correlation

Seungwook Kim Juhong Min Minsu Cho
Pohang University of Science and Technology (POSTECH), South Korea
http://cvlab.postech.ac.kr/research/HCCNet

Abstract

Recent studies show that leveraging the match-wise relationships within the 4D correlation map yields significant improvements in establishing semantic correspondences - but at the cost of increased computation and latency. In this work, we focus on the aspect that the performance improvements of recent methods can also largely be attributed to the usage of multi-scale correlation maps, which hold various information ranging from low-level geometric cues to high-level semantic contexts. To this end, we propose HCCNet, an efficient yet effective semantic matching method which exploits the full potential of multi-scale correlation maps, while eschewing the reliance on expensive match-wise relationship mining on the 4D correlation map. Specifically, HCCNet performs feature slicing on the bottleneck features to yield a richer set of intermediate features, which are used to construct a hypercolumn correlation. HCCNet can consequently establish semantic correspondences in an effective manner by reducing the volume of conventional high-dimensional convolution or self-attention operations to efficient point-wise convolutions. HCCNet demonstrates state-of-the-art or competitive performances on the standard benchmarks of semantic matching, while incurring a notably lower latency and computation overhead compared to the existing SoTA methods.

1. Introduction

Semantic correspondence is the task of establishing correspondences between two images depicting different instances of the same semantic category. While visual correspondence itself is a fundamental computer vision task used for 3D reconstruction, visual localization and object recognition [10], semantic correspondence has enabled further diverse applications, including semantic label/edit transfer [37, 40], unsupervised object discovery/localization [3], and few-shot classification/segmentation [14, 19, 20, 32]. While the recent success of deep neural networks in keypoint detection [1, 6] and feature descriptor extraction [41, 45] have shown significant improvements, the task of semantic correspondence remains challenging due to the presence of intra-class variations [4, 18, 21, 29, 31, 36, 44].

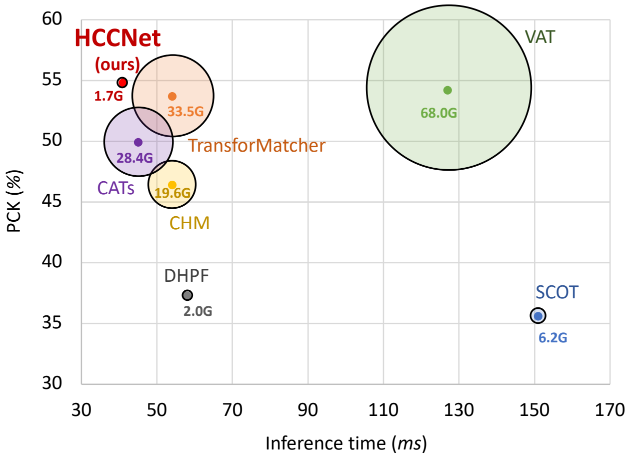

Figure 1. **PCK performance (y-axis) *vs.* inference time (x-axis) on SPair-71k dataset.** The area of each bubble is proportional to FLOPs of a model. We demonstrate that the proposed HCCNet outperforms existing state-of-the-art methods in terms of accuracy, efficiency, and scalability while being much simpler design than previous work [4, 14, 21, 29, 31, 36].

Among many effective learning-based methods that have been proposed by building on the efficacy of convolutional neural networks [17, 18, 34, 36, 42], a representative branch was largely inspired by the idea of learning geometric matching with high-dimensional convolution [25, 27, 31, 43], where convolutional layers are applied to the correlation map such that the certain unique matches would support the neighboring ambiguous matches. Noting that the convolution may suffer from inherent limitations of static and local transformations of the correlation map, current state-of-the-art methods propose to leverage self-attention to learn the global match-wise relations [4, 14, 21].

While leveraging the global interactions within a correlation map has shown to be highly effective, we suggest that the superior performance of today's state-of-the-art methods can also be attributed to the usage of multi-scale correlation maps (Section 4, Table 3). This is because semantic correspondences between images having large intra-class variations may occur at different feature levels, from local patterns and geometries (shallow) to invariant semantics and context (deep). It has also been demonstrated in other areas such as few-shot segmentation [14, 32] that leveraging multi-layer correlation maps shows improvements over using just a single correlation map.

In this work, we shift our focus away from mining global match-wise relations, to *better* leveraging the multi-scale correlation maps holding various semantics. To this end, we introduce an efficient yet effective semantic matching method, HCCNet, which carries out *feature slicing* to yield a richer set of equi-channel intermediate features from the backbone network, for constructing amplified multi-scale correlation maps. These multi-scale correlation maps are concatenated along the channel dimension to obtain a hypercolumn correlation. We finally perform a fast and efficient point-wise channel aggregation to output a refined correlation map for semantic keypoint transfer. The results demonstrate that our method surpasses existing state of the arts in terms of accuracy and efficiency despite its simple, straightforward design, as illustrated in Fig. 1.

The contributions of our work is threefold:

- We introduce HCCNet, a novel semantic matching learner that leverages various semantics of multi-scale correlations to establish reliable correspondences,
- We propose feature slicing, a method to yield rich feature slices from intermediate features to construct an informative hypercolumn correlation,
- The rich hypercolumn correlation enables HCCNet to reduce the volume of high-dimensional convolution or self-attention operations to point-wise channel aggregation, incurring notably lower computation and latency overhead while exhibiting SoTA or competitive performance on semantic correspondence.

2. Related Work

Leveraging multi-layer features and correlations for correspondence. For CNNs trained on the task of object recognition, the shallower layers learn geometric cues such as edge or color, and the deeper layers learn semantic cues of the object [11,12]. This characteristic of hierarchical features of CNNs have been applied to the task of establishing correspondences between images. Specifically, HPF [34] and its follow-up work, DHPF [36], propose to represent images using hyperpixels by leveraging a number of layers selected among early to late layers of the feature extractor. COLD [22] performs weighted summation on the intermediate feature maps to yield a distilled feature map pair. More recently, TransforMatcher [21] proposed to use multi-layer correlation maps, but only as features of each match position to be processed by match-to-match attention, without explicitly leveraging their consensus.

We focus on the aspect that semantic correspondences between images may occur at different feature levels depending on the image pair. To this end, we propose feature slicing to yield amplified multi-scale correlation maps to maximize the potential of the constructed hypercolumn correlation, on which we perform point-wise channel aggregation to exploit the various semantics of the correlation maps. We empirically demonstrate the superiority of our approach over concatenating or summing multi-level features to construct a single correlation map.

Consensus-based semantic correspondence. The task of semantic correspondence aims to establish correspondences between images of the same category but of different instances. While various CNN-based methods have been introduced to tackle this challenging problem [18, 25, 31, 36], with the recent advent of transformer-based architectures for visual tasks [7], transformer-based methods have demonstrated superior abilities to establish accurate semantic correspondences [4, 14, 21].

Among these approaches, NCNet [44] coined the idea of exploiting the local neighborhood consensus within the correlation map using high-dimensional convolutional networks. The efficacy of this approach motivated follow-up work to better exploit the neighborhood consensus to obtain reliable correspondences [25, 27, 31, 43] using high-dimensional CNNs. However, these methods suffer from the inherent limitations of CNNs *i.e*., local and static feature transformation. To alleviate these issues, the current SoTA methods on semantic correspondence exploit the *dynamic global* match-wise consensus in the correlation map [4, 14, 21] by building on self-attention mechanisms.

Albeit its efficacy, the endeavor to mine local or global match-wise relationships in the correlation map incurs high computation overhead. In this work, we propose to leverage hypercolumn correlation built from multi-scale correlation maps instead. As the multi-scale correlation maps are derived from feature maps of largely varying receptive fields, HCCNet implicitly considers the neighborhood consensus when performing feature matching after the point-wise channel aggregation despite its efficiency.

Attention for feature aggregation. Attention mechanisms enable neural networks to concentrate on the most relevant features, which has shown to be effective across many visual tasks such as object recognition and semantic segmentation [2, 15, 47–49]. SENet [15] exploits the channel-wise relationships by introducing the Squeeze-and-Excitation module. CBAM [48] combines the spatial and channel attention in a compact block. Bisenet [49] suggests a lightweight module for channel-wise attention. Noting the effectiveness of employing attention-based mech-

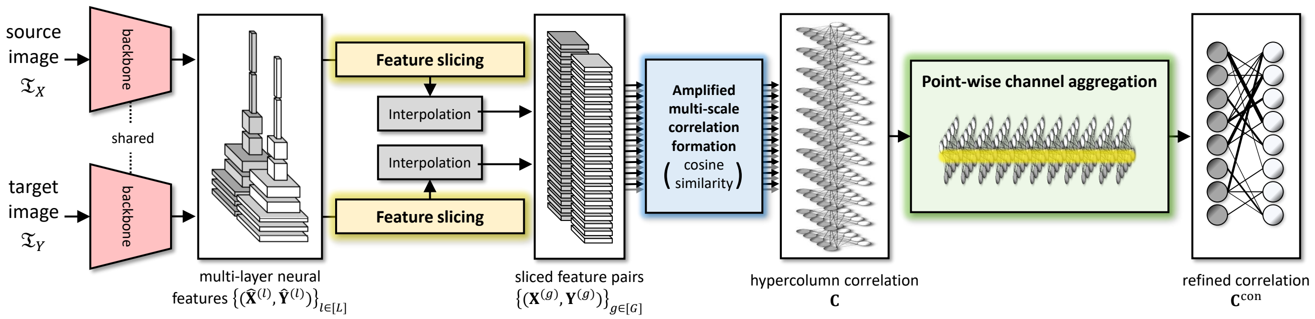
source image $\mathfrak{I}_X$
backbone
shared
backbone
target image $\mathfrak{I}_Y$
multi-layer neural features $\{(\hat{\mathbf{X}}^{(l)}, \hat{\mathbf{Y}}^{(l)})\}_{l\in[L]}$
Feature slicing
Interpolation
Interpolation
Feature slicing
sliced feature pairs $\{(\mathbf{X}^{(g)}, \mathbf{Y}^{(g)})\}_{g\in[G]}$
Amplified multi-scale correlation formation
(cosine similarity)
hypercolumn correlation
$\mathbf{C}$
Point-wise channel aggregation
refined correlation
$\mathbf{C}^{\mathrm{con}}$

Figure 2. **Overview of HCCNet.** The intermediate feature maps extracted from an image pair are first sliced, and are used to compute a consequently amplified multi-channel correlation map. We then identify and exploit the position-specific inter-correlation consensuses to provide the refined single-channel correlation map. We construct a dense flow field from the refined correlation map, which can be used to transfer given source keypoints to the target image to supervise HCCNet using ground-truth keypoint pair annotation.

anisms, more recent work propose to leverage attention to aggregate features. GFF [28] selectively fuses features from multiple levels using a gating mechanism in a fully connected manner. BPNet [38] uses an add-multiply-add fusion block to first add and multiply features from different levels separately, and then adds these two output features together. The weighted addition of features proposed in COLD [22] is also a form of attentive feature aggregation.

In this work, instead of fusing features extracted across the feature extractor, we propose to aggregate the channels of the hypercolumn *correlation* in a point-wise manner to obtain a refined correlation map for efficient and effective correspondence establishment.

3. HCCNet for semantic correspondence

We first provide an overview of how HCCNet establishes semantic correspondences. Given an image pair to match, we yield a set of intermediate feature maps using the backbone feature extractor network. These intermediate feature maps are bilinearly interpolated to the same spatial size, on which we perform *feature slicing* to yield a larger number of equi-channel feature maps. Each corresponding feature slice pair is used to calculate a single-channel correlation map, collectively yielding a set of multi-scale correlation maps. The multi-scale correlation maps are concatenated along the channel dimension to construct a *hypercolumn correlation*, on which we perform efficient point-wise convolution to aggregate the channels to output a refined correlation map. This refined correlation map is used to construct a dense flow field, which is used to transfer the given keypoints from the source images to the target image to establish correspondences between the image pair. Figure 2 illustrates the overall architecture of our method.

3.1. Feature slicing

We utilize an ImageNet-pretrained ResNet-101 [5, 13] as our feature extractor. To maximize the number of correlation maps and thus the visual cues to consider, we extract features from all bottleneck layers of `conv3_x`, `conv4_x`, and `conv5_x` blocks for a given pair of images $(\mathcal{I}_X, \mathcal{I}_Y)$. The multiple intermediate features are bilinearly interpolated to achieve the same (flattened) spatial dimension of HW; this dimension is $\frac{1}{16}$ of the input image resolution, thereby creating a set of features $\{(\hat{\mathbf{X}}^{(l)}, \hat{\mathbf{Y}}^{(l)})\}_{l\in[L]}$ where $\hat{\mathbf{X}}^{(l)}, \hat{\mathbf{Y}}^{(l)} \in \mathbb{R}^{HW\times C^{(l)}}$ represent the feature pair at layer l, $[L] := \{i\}_{i=1}^{L}$ represents a set of bottleneck layer indices, and $C^{(l)}$ indicates the channel size at layer l.

Previous related studies [4, 31] directly compute cosine similarity on extracted intermediate backbone feature pairs to form correlations, *i.e.*, $\hat{\mathbf{X}}^{(l)} \cdot \hat{\mathbf{Y}}^{(l)\top}$. However, such an approach could overlook *rich channel-wise information* of *high-dimensional backbone feature vectors* which potentially helps form richer correlation maps for the model to analyze. To address this issue, we introduce *feature slicing*, which slices each intermediate feature map, $\hat{\mathbf{X}}^{(l)}$ or $\hat{\mathbf{Y}}^{(l)}$, into multiple slices to provide a larger number of feature pairs by increasing the number of features to match. Specifically, we view each feature map at every layer as a composition of multiple sub-features concatenated along the channel dimension: $\hat{\mathbf{X}}^{(l)} := \text{concat}_{g\in G^{(l)}}\left[\mathbf{X}^{(g)}\right]$ for all $l \in [L]$, where $G^{(l)}$ is the number of slices used to divide feature map $\mathbf{X}^{(l)}$. This interpretation provides us with a more diverse set of visual features for the subsequent matching network to establish reliable matches, which we denote as $\{(\mathbf{X}^{(g)}, \mathbf{Y}^{(g)})\}_{g\in[G]}$ where $L < G$.

3.2. Hypercolumn correlation construction

To establish dense input pair-wise matches, we first calculate the cosine similarity between every possible position pairs between the feature maps. Specifically, for each group

$g \in [G]$, we compute the dense, richer ($L < G$) correlation matrix $\mathbf{C}_{:,:,g} \in \mathbb{R}^{HW \times HW}$ as follows:

$$\mathbf{C}_{\mathbf{x},\mathbf{y},g} = \frac{\mathbf{X}^{(g)}_{\mathbf{x},:} \cdot \mathbf{Y}^{(g)\top}_{\mathbf{y},:}}{\|\mathbf{X}^{(g)}_{\mathbf{x},:}\|_2 \|\mathbf{Y}^{(g)}_{\mathbf{y},:}\|_2}, \tag{1}$$

where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$ refer to the 2-dimensional spatial positions of the feature maps corresponding to the image pair of $\mathcal{I}_X$ and $\mathcal{I}_Y$ respectively. While some existing methods [21, 31] apply ReLU on top of correlation maps for non-negativity, we propose that the negative correlation scores also provide important cues for reliable correspondences, as empirically evidenced by better performance. After calculating the correlation maps for G feature slice pairs, we stack them along the channel dimension to produce the final hypercolumn correlation, denoted by $\mathbf{C} \in \mathbb{R}^{HW \times HW \times G}$. This approach enables us to consider a diverse set of intermediate feature pairs and provides richer information for the subsequent matching network to establish reliable matches.

3.3. Point-wise channel aggregation

When convolutional neural networks are trained on the task of object recognition, the feature representations become increasingly explicit about the object information along the processing hierarchy [11, 12]. Specifically, low-layer features contain more detailed information such as edge or colour, while higher-layer features contain more semantic information with higher invariance [8, 50]. Pertaining to the task of semantic matching, it is unsure at which feature layer the correlation is likely to be the most accurate, as the images depict different instances of the same class. Therefore, depending on the content of the given image pair, it may be beneficial to rely more on the lower-layer features, or rather on the higher-layer features. Now that we have a hypercolumn correlation $\mathbf{C}$ which holds the correlation information obtained from different layers of the backbone feature extractor, we aim to analyze the channels in a point-wise manner to aggregate the channels in order to yield the final refined correlation matrix.

It is crucial to ensure that a flow field has reliable and consistent information after analyzing different visual cues to establish reliable correspondences. To facilitate this, we aim to analyze the channels of the hypercolumn correlation for each spatial position $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^4$ in $\mathbf{C}$ to focus on or to downweight certain visual cues in aggregating the channels. Therefore, we collect match scores from different visual aspects of geometric and semantic cues and perform point-wise convolution as follows:

$$\Phi(\mathbf{C}; \mathbf{W}_{\text{hid}})_{\mathbf{x},\mathbf{y},:} \coloneqq \mathbf{C}_{\mathbf{x},\mathbf{y},:}\mathbf{W}_{\text{hid}} \in \mathbb{R}^{D_{\text{hid}}}, \tag{2}$$

where $\mathbf{W}_{\text{hid}} \in \mathbb{R}^{G \times D_{\text{hid}}}$ is a learnable weight matrix. To enhance the correlation consensus with better representational power, we process the correlation map $\mathbf{C}$ using two correlation consensus networks together with an intermediate hyperbolic tangent non-linearity ζ:

$$(\mathbf{C}^{\text{con}})_{\mathbf{x},\mathbf{y}} \coloneqq \Phi(\zeta(\Phi(\mathbf{C}; \mathbf{W}_{\text{hid})}))); \mathbf{W}_{\text{out}})_{\mathbf{x},\mathbf{y}} \tag{3}$$

$$= \zeta(\mathbf{C}_{\mathbf{x},\mathbf{y},:}\mathbf{W}_{\text{hid})})\mathbf{W}_{\text{out}} \in \mathbb{R}, \tag{4}$$

where $\mathbf{W}_{\text{out}} \in \mathbb{R}^{D_{\text{hid}} \times 1}$ is learnable matrix that squeezes multiple channels to provide a single, refined correlation map for the subsequent flow field formation.

Note that we employ a simple yet effective approach that leverages hypercolumn correlation, striking a balance between efficacy and efficiency without resorting to overly complex methodologies *e.g.* match-wise relation mining, for visual correspondence. In Section 4, we present empirical evidence that highlights the efficacy of our method despite its straightforward nature, showing that it surpasses existing methods without relying on computationally-demanding techniques, *e.g.*, high-dimensional convolutions [31, 33], cost aggregations [4], or Hough matching [34, 36]. By avoiding such complicated methodologies and instead relying on straightforward, simple design by leveraging pretrained backbone features, we pave the way for more accessible, scalable, and practical solutions to the problem of visual correspondence.

3.4. Flow field formation and keypoint transfer

For fine-grained flow field formation, the aggregated correlation matrix $\mathbf{C}^{\text{con}}$ is then upsampled via a 4-dimensional upsampling function that provides $\mathbf{C}^{\text{out}} \in \mathbb{R}^{\bar{H}\bar{W} \times \bar{H}\bar{W}}$ where $\bar{H} = 4H$ and $\bar{W} = 4W$, which corresponds to $\frac{1}{4}$ the size of the original image. We use the output correlation tensor $\mathbf{C}^{\text{out}}$ to form a dense flow field between the source and target image for keypoint transfer. First, we normalize the the output correlation map by applying kernel soft-argmax [23] as follows:

$$\mathbf{C}^{\text{norm}}_{\mathbf{x},\mathbf{y}} = \frac{\exp(\mathbf{G}^{\mathbf{p}}_{\mathbf{y}}\mathbf{C}^{\text{out}}_{\mathbf{x},\mathbf{y}})}{\sum_{\mathbf{m} \in [\bar{H}] \times [\bar{W}]} \exp(\mathbf{G}^{\mathbf{p}}_{\mathbf{m}}\mathbf{C}^{\text{out}}_{\mathbf{x},\mathbf{m}})} \in \mathbb{R}, \tag{5}$$

where $\mathbf{G}^{\mathbf{p}} \in \mathbb{R}^{\bar{H} \times \bar{W}}$ is a 2D Gaussian kernel centered on $\mathbf{p} = \arg\max_{\mathbf{y}} \mathbf{C}^{\text{out}}_{\mathbf{x},\mathbf{y}}$, to suppress noisy correlation values in the correlation map. The normalized correlation tensor $\mathbf{C}^{\text{norm}}$ encodes a set of probability simplexes from each source feature position to the target feature positions. We then transfer all the coordinates on the dense regular grid $\mathbf{P}_X \in \mathbb{R}^{\bar{H}\bar{W} \times 2}$ of source image $\mathcal{I}_X$ to obtain their corresponding coordinates $\hat{\mathbf{P}}_Y \in \mathbb{R}^{\bar{H}\bar{W} \times 2}$ on target image $\mathcal{I}_Y$:

$$(\hat{\mathbf{P}}_Y)_{\mathbf{x},:} = \sum_{(\mathbf{y}) \in [\bar{H}] \times [\bar{W}]} \mathbf{C}^{\text{norm}}_{\mathbf{x},\mathbf{y}}(\mathbf{P}_X)_{\mathbf{y},:} \in \mathbb{R}^2, \tag{6}$$

forming a dense flow field. Using this dense flow field, we can perform keypoint transfer as follows. Given a keypoint

$\mathbf{k}^X = (x_k, y_k)$, we define a soft sampler $\mathbf{W}^{(k)} \in \mathbb{R}^{\bar{H}\times\bar{W}}$:

$$\mathbf{W}^{(k)}_{ij} = \frac{\max(0, \tau - \sqrt{(x_k - j)^2 + (y_k - i)^2})}{\sum_{i'j'} \max(0, \tau - \sqrt{(x_k - j')^2 + (y_k - i')^2})}, \tag{7}$$

where τ is a distance threshold, and $\sum_{ij} \mathbf{W}^{(k)}_{ij} = 1$. The above equation shows that the soft sampler samples each transferred keypoint $(\hat{\mathbf{P}}_Y)_{ij}$ by assigning weights which are inversely proportional to the distance to $\mathbf{k}^X$. Using this soft sampler, we assign a match to the keypoint $\mathbf{k}^X$ as $\mathbf{k}^Y = \sum_{(i,j)\in[\bar{H}]\times[\bar{W}]} (\hat{\mathbf{P}}_Y)_{ij:} \mathbf{W}^{(k)}_{ij}$, being able to establish sub-pixel-wise accurate correspondences.

3.5. Training objective

For each training image pair with ground-truth correspondences $\mathcal{M} = \{(\hat{\mathbf{k}}^X_m, \hat{\mathbf{k}}^Y_m)\}^M_{m=1}$, we apply the aforementioned keypoint transfer method on the given source keypoints to obtain predicted target keypoints. This results in a set of predicted correspondences $\{(\hat{\mathbf{k}}^X_m, \mathbf{k}^Y_m)\}^M_{m=1}$ by assigning a match $\mathbf{k}^Y_m$ to each keypoint $\hat{\mathbf{k}}^X_m$ in the source image. We formulate our training objective to minimize the average Euclidean distance between the predicted target keypoints and the ground-truth target keypoints as follows:

$$\mathcal{L} = \frac{1}{M} \sum_{m=1}^{M} \|\mathbf{k}^Y_m - \hat{\mathbf{k}}^Y_m\|^2_2. \tag{8}$$

4. Experiments

In this section, we evaluate HCCNet against the state-of-the-art methods on the task of semantic matching and discuss the results with in-depth analysis.

4.1. Evaluation settings

Implementation details. We use the ImageNet [5]-pretrained ResNet-101 model [13] as our feature extractor. The `conv3_x`, `conv4_x` and `conv5_x` layers have 4, 23 and 3 bottleneck layers, respectively; we utilize all these bottleneck layers, and use feature slices with channel dimension of 256 to finally yield $G = 124$ feature slice pairs to construct a 124-layer correlation map for an input image pair ($G = 30$). We use an image size of 240×240 for both training and inference, where the feature map dimensions used for correlation computation is $H = W = 15$, and the upsampled $\mathbf{C}^{\text{out}}$ has spatial dimensions of $\bar{H} = \bar{W} = 60$. Both the number of groups and channel size of linear layers in correlation consensus network are set to 124, *i.e.*, $G = D_{\text{hid}} = 124$. HCCNet is implemented using PyTorch [39], and our network is optimized using the AdamW [30] optimizer with a learning rate of 1e-3 for the correlation network, and 1e-5 for the ResNet-101 feature extractor.

Datasets. We evaluate our method on the standard benchmark datasets of semantic matching: PF-PASCAL, PF-WILLOW [9] and SPair-71k [35] with keypoint-annotated image pairs. PF-PASCAL consists of image pairs from the PASCAL VOC 2007 dataset, having the same viewpoint and small scale variations. PF-PASCAL contains 2,940 / 308 / 299 image pairs for training, validation and testing, respectively. PF-WILLOW is comprised of four categories of the PASCAL VOC 2007 and Caltech-256 datasets, having center-aligned image pairs with the same viewpoint and small scale variations. PF-WILLOW contains 900 image pairs for testing only. SPair-71k consists of image pairs from PASCAL3D+, and PASCAL VOC 2012 datasets, with diverse variations in viewpoint and scale. SPair-71k has 53,340 / 5,384 / 12,234 image pairs for training, validation, and testing, respectively. The results on SPair-71k are much less saturated in comparison to other benchmarks due to its large scale and challenging variations.

Evaluation metric. We use the percentage of correct keypoints (PCK) as the evaluation metric. Given a pair of ground-truth keypoints and our predicted target keypoints, the PCK can be computed as follows:

$$\text{PCK}(\mathcal{K}) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}[\|\mathbf{k}^Y_m - \hat{\mathbf{k}}^Y_m\| \leq \alpha_\tau \cdot \max(w_\tau, h_\tau)], \tag{9}$$

where w_τ and h_τ denotes the width and height thresholds, which are the width and height of either the entire image or the object bounding box, *i.e.*, $\tau \in \{\text{img}, \text{bbox-kp}, \text{bbox}\}$, and α_τ is a tolerance factor.

4.2. Quantitative results on semantic matching

Table 1 illustrates the quantitative results of HCCNet in comparison to existing methods on the standard benchmarks of semantic matching. To directly demonstrate the efficacy of our method, we report finetuned (F) results, which are trained on the train set of the corresponding dataset. To evaluate the cross-dataset generalizability, we report transferred (T) results, where we use a model trained on the train set of PF-PASCAL for evaluation. It can be seen that HCCNet sets a new state of the art on the finetuned (F) setting of the SPair-71k dataset, which is the most challenging semantic matching benchmark, while being competitive on the finetuned (F) setting of the PF-PASCAL dataset, just 0.2%p below CATs† [4]. It is noteworthy that HCCNet achieves this while incurring notably lower latency and FLOPs compared to existing methods. On the contrary, HCCNet yields subpar outcomes on the transferred (T) settings, which we conjecture is due to HCCNet's brittleness to the domain gap between datasets, resulting in inconsistent point-wise channel aggregation. The classwise PCK results on SPair-71k is shown in Table 2, and Figure 3 visualizes

Method	SPair-71k		PF-PASCAL		PF-WILLOW		time	memory	FLOPs
	@α_{bbox}		@α_{img}		@$\alpha_{\text{bbox-kp}}$	@α_{bbox}	(*ms*)	(GB)	(G)
	0.1 (F)	0.1 (T)	0.05 (F)	0.1 (F)	0.1 (T)	0.1 (T)			
HPF [34]	28.2	-	60.1	84.8	74.4	-	63	-	-
SCOT [29]	35.6	-	63.1	85.4	**76.0**	-	151	4.6	6.2
DHPF [36]	37.3	27.4	75.7	90.7	71.0	77.6	58	1.6	2.0
DHPF† [36]	39.4	-	-	-	-	-	58	1.6	2.0
NC-Net* [44]	-	-	-	81.9	-	-	222	1.2	44.9
DCC-Net* [16]	-	-	-	83.7	-	-	567	2.7	47.1
ANC-Net [25]	-	28.7	-	86.1	-	-	216	**0.9**	44.9
PMD [26]	37.4	-	-	90.7	75.6	-	-	-	-
CHMNet [31]	46.4	30.1	80.1	91.6	69.6	79.4	54	1.6	19.6
PMNC [24]	50.4	-	**82.4**	90.6	-	-	-	-	-
MMNet [51]	40.9	-	77.6	89.1	-	-	86	-	-
CATs [4]	43.5	-	-	-	-	-	45	1.6	28.4
CATs† [4]	49.9	27.1	75.4	**92.6**	69.0	79.2	45	1.6	28.4
PWarpC-NC-Net [46]	52.0	**37.1**	67.8	82.3	-	76.2	-	-	-
TransforMatcher [21]	50.2	30.5	78.9	90.5	66.7	75.1	54	1.6	33.5
TransforMatcher† [21]	53.7	30.1	80.8	91.8	65.3	76.0	54	1.6	33.5
VAT† [14]	54.2	-	78.2	92.3	-	**81.6**	127	3.6	68.0
HCCNet (ours)	53.9	29.6	80.2	92.4	65.3	74.5	**30**	2.0	**1.7**
HCCNet † (ours)	**54.8**	29.7	80.2	92.4	65.5	74.5	**30**	2.0	**1.7**

Table 1. **Performance on standard benchmarks of semantic matching.** All the methods reported in the above table uses a pretrained ResNet-101 model as the feature extractor. The first group of methods were trained with weak supervision (image pair annotations), and the second group of methods were trained using strong supervision (keypoint pair annotations). Models with * are retrained using keypoint annotations from ANC-Net [25]. † indicates the use of data augmentation during training. Numbers in bold indicate the best performance, followed by the underlined numbers.

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	dog	horse	mbike	person	plant	sheep	train	tv	all
NC-Net [44]	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
HPF [34]	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	14.9	31.5	35.6	28.2
SCOT [29]	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 [36]	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
CHMNet [31]	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	44.2	38.9	31.3	72.2	55.6	46.4
PMNC [24]	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
MMNet [51]	43.5	27.0	62.4	27.3	40.1	50.1	37.5	60.0	21.0	56.3	50.3	41.3	30.9	19.2	30.1	33.2	64.2	43.6	40.9
CATs [4]	46.5	26.9	69.1	24.3	44.3	38.5	30.2	65.7	15.9	53.7	52.2	46.7	32.7	35.2	32.2	31.2	68.0	49.1	43.5
CATs† [4]	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
TransforMatcher [21]	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
TransforMatcher [21]†	59.2	39.3	73.0	41.2	52.5	**66.3**	55.4	67.1	26.1	67.1	56.6	53.2	45.0	39.9	42.1	35.3	75.2	68.6	53.7
VAT† [14]	56.5	37.8	73.0	38.7	50.9	58.2	40.9	70.5	20.3	**72.1**	**61.1**	57.7	45.6	48.2	**52.4**	**40.0**	77.7	**71.4**	54.2
HCCNet	**59.9**	39.1	71.0	**42.1**	51.6	63.4	**57.0**	63.0	**26.8**	63.8	59.4	54.7	**49.4**	41.0	43.0	37.6	**83.1**	64.8	53.9
HCCNet †	**59.9**	**40.6**	70.5	39.8	**55.9**	65.1	56.8	66.6	25.6	69.2	59.6	**58.7**	46.7	40.3	43.6	39.6	82.2	65.4	**54.8**

Table 2. **Classwise PCK on SPair-71k.** All the methods reported in the above table uses a pretrained ResNet-101 model as the feature extractor. † indicates the use of data augmentation during training. We take results from the methods whose classwise PCK results were provided. Numbers in bold indicate the best performance, followed by the underlined numbers.

example qualitative results on the test set of SPair-71K in comparison to TransforMatcher [21].

4.3. Ablation study and analysis

Effect of using multiple correlation maps in existing methods. Table 3 illustrates the performance of existing methods when using a single correlation map in comparison to using multiple correlation maps. It is visible that the significant gain in performance is consistent across different methods, substantiating our claim that the efficacy of today's SoTA methods can be largely attributed to the usage of muliple correlation maps[1].

[1] While TransforMatcher$_{\text{mean}}$ or TransforMatcher$_{\text{concat}}$ use multi-level features, they yield a single correlation map as a result of mean or concate-

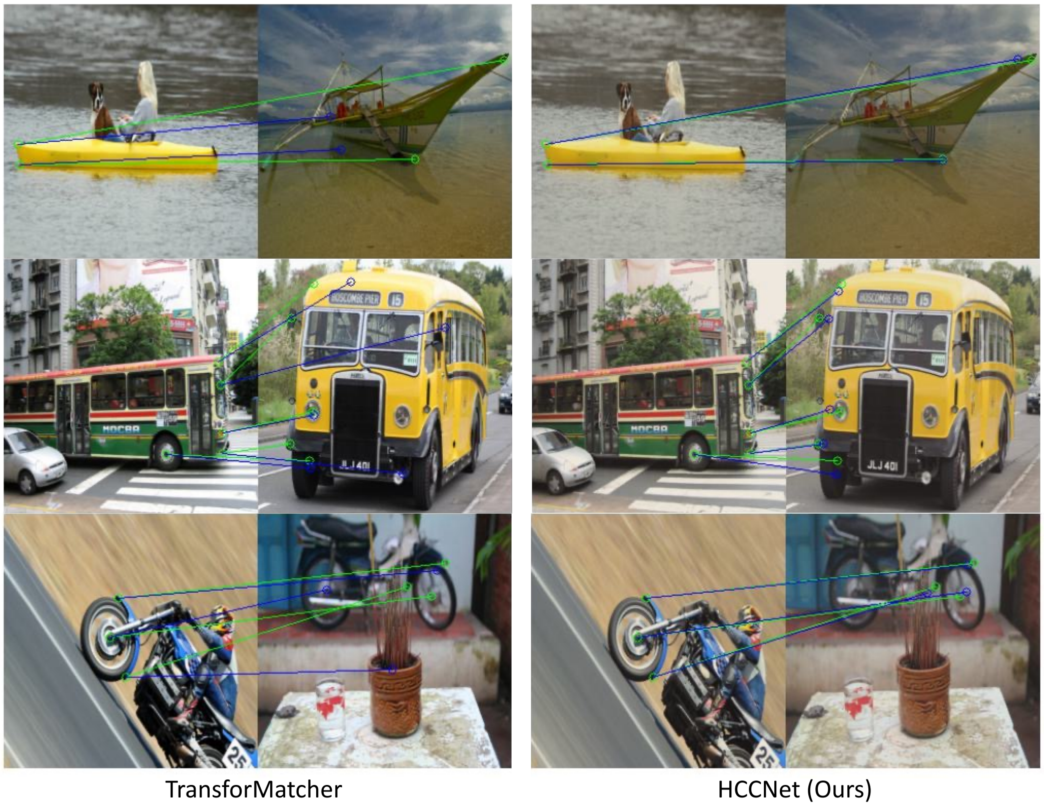

Figure 3. **Qualitative comparison of HCCNet against TransforMatcher** [21]. Green lines represent ground truth correspondences, and blue lines represent predicted correspondences. Best viewed on electronics.

Method	SPair-71k (α_{img})	
	0.05	0.1
CHMNet$_{\text{conv3\_x}}$ [31]	-	47.0
CHMNet$_{\text{multi}}$ [33]	-	51.3
*CATs$_{\text{conv3\_x}}$ [4]	26.2	48.3
CATs$_{\text{multi}}$ [4]	27.7	49.9
TransforMatcher$_{\text{concat}}$ [21]	20.9	41.7
TransforMatcher$_{\text{mean}}$ [21]	24.1	45.1
TransforMatcher$_{\text{multi}}$ [21]	32.4	53.7

Table 3. **PCK performance of existing methods when using a single correlation map v.s. multiple correlation maps on the SPair-71k dataset.** The results are taken from their reported results, except for *CATs$_{\text{conv3\_x}}$ which was implemented by us.

`conv` used				PF-PASCAL @α_{img}		mem. (GB)	FLOPs (G)
2_x	3_x	4_x	5_x	0.05	0.1		
×	×	×	✓	72.1	90.1	1.9	0.9
×	×	✓	×	79.5	91.6	1.9	1.3
×	×	✓	✓	80.2	91.8	1.9	1.6
×	✓	✓	✓	80.2	92.4	2.0	1.7
✓	✓	✓	✓	79.9	92.1	2.0	1.7

Table 4. **Ablation study on the backbone bottleneck features used.** The results show that our current setting of using `conv3_x` to `conv5_x` yields the best results.

Ablation study on the backbone convolutional blocks used. We compare the results of HCCNet when extracting bottleneck features from varying convolutional blocks of the backbone network. The results in Table 4 shows that our current setting of using `conv3_x` to `conv5_x` strikes the best balance between performance and efficiency.

Analysis on the feature slice size. We compare the results of HCCNet when using varying sizes of feature slices, or when directly using the bottleneck features for correlation computation. The results in Table 5 show that our current setting of using a slice size of 256 yields a favorable balance between performance and efficiency, and the latency and FLOPs increases dramatically with decreasing slice size.

Analysis on the non-linear activation function used. Table 6 shows that using the hyperbolic tangent (Tanh) non-linear activation function yields favorable results in comparison to ReLU or Sigmoid functions. We conjecture this is

nation of the multi-level features to yield a single feature map pair.

Slice size	PF-PASCAL $@\alpha_{\text{img}}$ 0.05	 0.1	time (*ms*)	mem. (GB)	FLOPs (G)
-	77.3	92.2	20	2.0	0.9
512	77.0	91.9	24	2.0	1.1
256	80.2	92.4	30	2.0	1.7
128	80.2	92.2	43	2.2	4.0
64	79.5	92.5	70	2.2	13.3
32	80.4	92.4	127	2.7	50.7
16	79.9	91.5	290	3.8	200
8	65.0	82.5	580	5.0	798

Table 5. **Ablation study on the slice size used.** The results show that our current setting of using the chunk size of 256 yields the best trade-off between performance and efficiency.

Activation function	PF-PASCAL $@\alpha_{\text{img}}$ 0.05	 0.1
ReLU	79.5	91.9
Sigmoid	79.6	91.8
Tanh	**80.2**	**92.4**

Table 6. **Ablation study on the non-linear activation function used.** Using the Tanh activation function yields the best results, over ReLU or Sigmoid activation functions.

because unlike ReLU or Sigmoid, Tanh is capable of representing unlikely matches using negative correlation scores.

Feature slicing analysis. To investigate the impact of channel aggregation on the hypercolumn correlation, we visualize learned weight matrices $\mathbf{W}_{\text{hid}}$ and $\mathbf{W}_{\text{out}}$ with four different groups denoted by $G \in \{30, 62, 124, 248\}$[2] in Fig. 4. We observe that the weight magnitudes are notably higher (in yellow) at deeper layers, particularly at `conv4_x` and `conv5_x`, as opposed to shallower layers. As we increase the number of groups utilized for feature slicing, we find that the network carries out *fine-grained channel selection*, as evidenced by the weight visualization of $\mathbf{W}_{\text{hid}}$, verifying the efficacy of performing position-wise channel aggregation on hypercolumn correlation using diverse visual cues. Compared to the weight magnitudes of $\mathbf{W}_{\text{hid}}$ that are focused on specific groups, those of $\mathbf{W}_{\text{out}}$ are relatively evenly dispersed in order to effectively aggregate the information from the first channel aggregation to provide a reliable refined correlation map.

We guide the readers to the supplementary for more analyses and experiments of HCCNet.

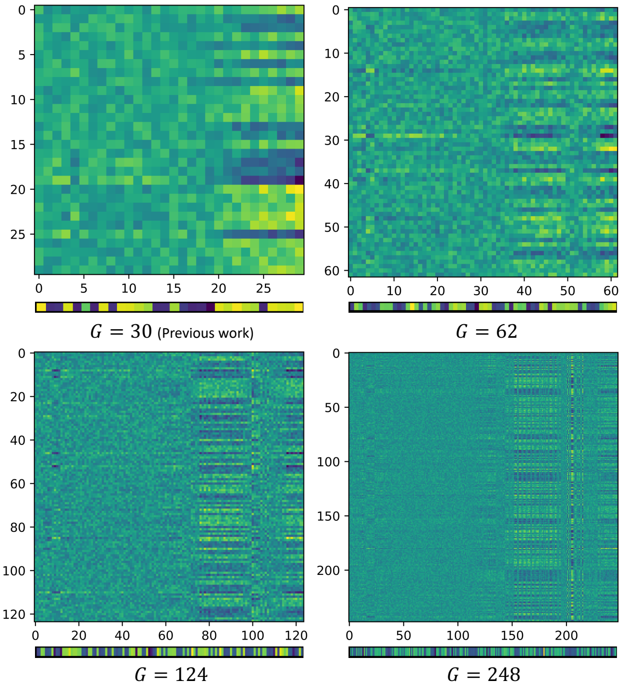

Figure 4. Visualization of learned weight matrices of $\mathbf{W}_{\text{hid}} \in \mathbb{R}^{G \times D_{\text{hid}}}$ (top) and $\mathbf{W}_{\text{out}} \in \mathbb{R}^{D_{\text{hid}} \times 1}$ (bottom) under varying $G = D_{\text{hid}} \in \{30, 62, 124, 248\}$.

5. Conclusion

In this work, we introduced HCCNet, an efficient yet effective method to establish semantic correspondences between images. Noting that the current trend of mining inter-match relations within the correlation map is computationally demanding, we shifted our focus to *better* leveraging the multi-level correlation maps computed from feature maps of varying receptive fields and visual cues. Our technical edge lies in the synergistic integration of our proposed feature slicing and point-wise convolution; by leveraging feature slicing to yield a richer set of intermediate features, HCCNet can effectively establish semantic correspondences while reducing the volume of conventional high-dimensional convolution operations to point-wise convolutions. Attributing to the eschewal of match-wise relation mining on the correlation map, HCCNet incurs notably lower latency and computation overhead while achieving state-of-the-art or competitive performance on the standard benchmarks of semantic correspondence.

Acknowledgements. This work was supported by the NRF grant (NRF-2021R1A2C3012728 (50%)) and the IITP grants (2022-0-00290: Visual Intelligence for Space-Time Understanding and Generation based on Multi-layered Visual Common Sense (40%), 2019-0-01906: AI Graduate School Program at POSTECH (10%)) funded by Ministry of Science and ICT, Korea.

[2] Note that using $G = 30$ means that feature slicing is not performed, as the total number of intermediate features extracted across the bottleneck layers is already 30.

References

[1] Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key.Net: Keypoint Detection by Handcrafted and Learned CNN Filters. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*, 2019. 1

[2] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016. 2

[3] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *CVPR*, 2015. 1

[4] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Cats: Cost aggregation transformers for visual correspondence. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 1, 2, 3, 4, 5, 6, 7

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3, 5

[6] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR Deep Learning for Visual SLAM Workshop*, 2018. 1

[7] 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. *ICLR*, 2021. 2

[8] Chen Du, Yanna Wang, Chunheng Wang, Cunzhao Shi, and Baihua Xiao. Selective feature connection mechanism: Concatenating multi-layer cnn features with a feature selector. *Pattern Recognition Letters*, 129:108–114, 2020. 4

[9] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, Jan 2015. 5

[10] David Forsyth and Jean Ponce. *Computer Vision: A Modern Approach. (Second edition)*. Prentice Hall, Nov. 2011. 1

[11] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *NIPS*, 2015. 2, 4

[12] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016. 2, 4

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 5

[14] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2022. 1, 2, 6

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 2

[16] Shuaiyi Huang, Qiuyue Wang, Songyang Zhang, Shipeng Yan, and Xuming He. Dynamic context correspondence network for semantic alignment. In *ICCV*, 2019. 6

[17] Sangryul Jeon, Seungryong Kim, Dongbo Min, and Kwanghoon Sohn. Parn: Pyramidal affine regression networks for dense semantic correspondence. In *ECCV*, 2018. 1

[18] Sangryul Jeon, Dongbo Min, Seungryong Kim, Jihwan Choe, and Kwanghoon Sohn. Guided semantic flow. In *ECCV*, 2020. 1, 2

[19] Dahyun Kang and Minsu Cho. Integrative few-shot learning for classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9979–9990, June 2022. 1

[20] Dahyun Kang, Heeseung Kwon, Juhong Min, and Minsu Cho. Relational embedding for few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8822–8833, October 2021. 1

[21] Seungwook Kim, Juhong Min, and Minsu Cho. Transformatcher: Match-to-match attention for semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8697–8707, 2022. 1, 2, 4, 6, 7

[22] Jongmin Lee, Yoonwoo Jeong, Seungwook Kim, Juhong Min, and Minsu Cho. Learning to distill convolutional features into compact local descriptors. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 897–907, 2021. 2, 3

[23] Junghyup Lee, Dohyung Kim, Jean Ponce, and Bumsub Ham. Sfnet: Learning object-aware semantic correspondence. In *CVPR*, 2019. 4

[24] Jae Yong Lee, Joseph DeGol, Victor Fragoso, and Sudipta Sinha. Patchmatch-based neighborhood consensus for semantic correspondence. In *CVPR*, 2021. 6

[25] Shuda Li, Kai Han, Theo W. Costain, Henry Howard-Jenkins, and Victor Prisacariu. Correspondence networks with adaptive neighbourhood consensus. In *CVPR*, 2020. 1, 2, 6

[26] Xin Li, Deng-Ping Fan, Fan Yang, Ao Luo, Hong Cheng, and Zicheng Liu. Probabilistic model distillation for semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7505–7514, June 2021. 6

[27] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-resolution correspondence networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2

[28] Xiangtai Li, Houlong Zhao, Lei Han, Yunhai Tong, Shaohua Tan, and Kuiyuan Yang. Gated fully fusion for semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11418–11425, 2020. 3

[29] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *CVPR*, 2020. 1, 6

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[31] Juhong Min and Minsu Cho. Convolutional hough matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2940–2950, June 2021. 1, 2, 3, 4, 6, 7

[32] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[33] Juhong Min, SeungWook Kim, and Minsu Cho. Convolutional hough matching networks for robust and efficient visual correspondence. *arXiv preprint arXiv:2109.05221*, 2021. 4, 7

[34] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *ICCV*, 2019. 1, 2, 4, 6

[35] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. SPair-71k: A large-scale benchmark for semantic correspondence. *arXiv prepreint arXiv:1908.10543*, 2019. 5

[36] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Learning to compose hypercolumns for visual correspondence. In *ECCV*, 2020. 1, 2, 4, 6

[37] Jiteng Mu, Shalini De Mello, Zhiding Yu, Nuno Vasconcelos, Xiaolong Wang, Jan Kautz, and Sifei Liu. Coordgan: Self-supervised dense correspondences emerge from gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10011–10020, June 2022. 1

[38] Dong Nie, Jia Xue, and Xiaofeng Ren. Bidirectional pyramid networks for semantic segmentation. In *Proceedings of the asian conference on computer vision*, 2020. 3

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[40] William Peebles, Jun-Yan Zhu, Richard Zhang, Antonio Torralba, Alexei Efros, and Eli Shechtman. Gan-supervised dense visual alignment. In *CVPR*, 2022. 1

[41] Jerome Revaud, Philippe Weinzaepfel, César Roberto de Souza, and Martin Humenberger. R2D2: repeatable and reliable detector and descriptor. In *NeurIPS*, 2019. 1

[42] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. End-to-end weakly-supervised semantic alignment. In *CVPR*, 2018. 1

[43] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Efficient neighbourhood consensus networks via submanifold sparse convolutions. In *ECCV*, 2020. 1, 2

[44] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *NeurIPS*, 2018. 1, 2, 6

[45] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. Sosnet: Second order similarity regularization for local descriptor learning. In *CVPR*, pages 11016–11025, 2019. 1

[46] Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Probabilistic warp consistency for weakly-supervised semantic correspondences. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8698–8708, 2022. 6

[47] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017. 2

[48] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 2

[49] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 2

[50] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013. 4

[51] Dongyang Zhao, Ziyang Song, Zhenghao Ji, Gangming Zhao, Weifeng Ge, and Yizhou Yu. Multi-scale matching networks for semantic correspondence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 6