JOINT LEARNING OF FEATURE EXTRACTION AND COST AGGREGATION FOR SEMANTIC CORRESPONDENCE

Jiwon Kim, Youngjo Min, Mira Kim, and Seungryong Kim

Computer Vision Lab. (CVLAB), Korea University, Korea

ABSTRACT

Establishing dense correspondences across semantically similar images is one of the challenging tasks due to the significant intra-class variations and background clutters. To solve these problems, numerous methods have been proposed, focused on learning feature extractor or cost aggregation independently, which yields sub-optimal performance. In this paper, we propose a novel framework for jointly learning feature extraction and cost aggregation for semantic correspondence. By exploiting the pseudo labels from each module, the networks consisting of feature extraction and cost aggregation modules are simultaneously learned in a boosting fashion. Moreover, to ignore unreliable pseudo labels, we present a confidence-aware contrastive loss function for learning the networks in a weakly-supervised manner. We demonstrate our competitive results on standard benchmarks for semantic correspondence.

Index Terms— semantic correspondence, feature extraction, cost aggregation, contrastive learning

1. INTRODUCTION

Semantic correspondence is one of the essential tasks on various Computer Vision applications [1, 2, 3], which generally aims to establish pixel-wise, but locally-consistent correspondences across semantically similar images. It is an extremely challenging task because finding semantic correspondences can be easily distracted by non-rigid deformations and large variations on the appearance within the same class [4].+-

Recent methods [5, 3] solved the task by designing deep Convolutional Neural Networks (CNNs). The networks often consist of *feature extraction* and *cost aggregation* steps. For the feature extraction step, instead of relying on handcrafted descriptors as in conventional methods [6], recently, there has been an increasing interest in leveraging the representation power of CNNs [7, 8, 4]. However, they can struggle with determining the correct matches from the cost volume because ambiguous matching pairs are often generated by repetitive patterns and occlusions. For cost aggregation step, methods [9, 10, 5, 11, 12, 13, 14, 3] attempted to determine the correct matches between the great majority of dense information and non-distinctive matching pairs. Unlike previous strategies [6, 15], recent methods proposed a trainable matching cost aggregation in the overall network [9, 10, 5, 11, 12, 13, 14, 3].

Learning semantic correspondence networks consisting of feature extraction and cost aggregation modules in a *supervised* manner requires a large-scale ground-truth which is notoriously hard to build. To alleviate this, several methods leveraged *pseudo-labels*, extracted from networks' prediction itself by Winner-Take-All (WTA), and train the networks in an unsupervised[1] manner [10, 5]. Although they are appealing alternatives, they are sensitive to uncertain pseudo labels. Furthermore, jointly using the pseudo labels from feature extraction and cost aggregation modules may boost the performance, but there was no study for this approach.

Some recent self-supervised methods [16, 17] use dense contrastive loss for pixel-wise prediction tasks instead of using image-level contrastive loss. However, they do not consider unconfident matches generated from repetitive fields or occlusions and semantic appearance variations.

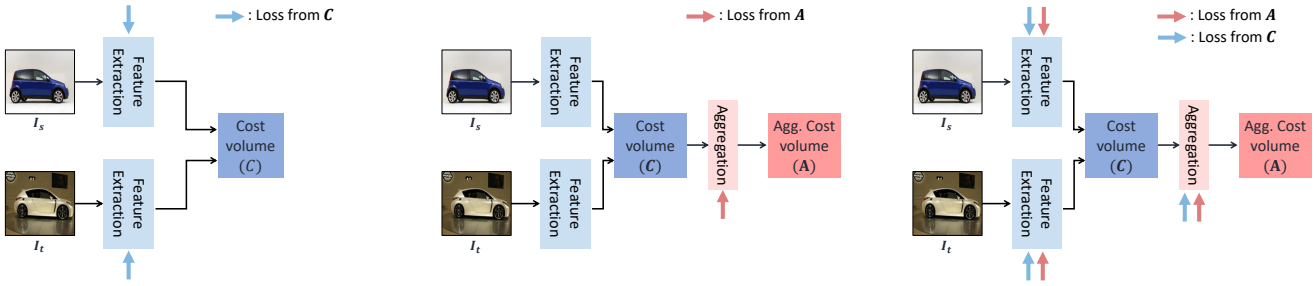
In this work, we present a novel framework that jointly learns feature extraction and cost aggregation for semantic correspondence. Motivated by the observation that the pseudo labels from feature extraction and cost aggregation steps can complement each other, we encourage the feature extraction and cost aggregation modules to be jointly trained by exploiting the pseudo-labels from each other module. In addition, to filter out unreliable pseudo-labels, we present a confidence-aware contrastive loss function that exploits a forward-backward consistency to remove incorrect matches. We demonstrate experiments on several benchmarks [18, 19], proving the robustness of the proposed method over the latest methods for semantic correspondence.

2. METHODOLOGY

2.1. Preliminaries

Given a pair of images, i.e., source I_s and target I_t, which represent semantically similar images, the objective of semantic correspondence is to predict dense correspondence fields P between the two images at each pixel. To achieve this, most predominant methods consist of two steps, namely *feature*

[1]They are often also called weak-supervised methods since they require the image pairs.



(a) Learning feature extraction only (b) Learning cost aggregation only (c) Ours

Fig. 1: **Intuition of our method:** (a) methods for solely training a feature extraction network [7, 8, 4], (b) methods for solely training a cost aggregation network for filtering the ambiguous matching cost [9, 10, 5, 11, 14], and (c) Ours, which jointly training feature extraction and cost aggregation networks with the proposed loss function.

extraction and *cost aggregation* [8, 5]. In specific, the first step is to apply feature extraction networks to obtain a 3D tensor $D \in \mathbb{R}^{h \times w \times d}$, where $h \times w$ is the spatial resolution of the feature maps and d denotes the number of channels. To estimate correspondences, the similarities between feature maps D_s and D_t from source and target images, respectively, are measured, which outputs a 4D cost volume $\mathcal{C}(i,j)$, where $i,j \in \{1,...,h \times w\}$, through a cosine similarity score such that $\mathcal{C}(i,j) = D_s(i)^T D_t(j)$. Estimating the correspondence with sole reliance on matching similarities is sensitive to matching outliers, and thus the cost aggregation steps are used to refine the initial matching similarities to achieve the aggregated cost $\mathcal{A}(i,j)$ through cost aggregation networks.

Learning such networks, i.e., feature extraction and cost aggregation modules, in a *supervised* manner requires manually annotated ground-truth correspondences P^*, which is extremely labor-intensive and subjective to build [19, 18, 5]. To overcome this challenge, an alternative way is to leverage Winner-Take-All (WTA) matching point, which is the most likely match by an argmax function on $\mathcal{C}(i,j)$ or $\mathcal{A}(i,j)$, as a pseudo correspondence label F. For instance, NCNet [5] (and its variants [13, 12, 14]) and DenseCL [17] utilized such correspondences F to learn the cost aggregation networks and feature extraction networks, respectively, in an *unsupervised* fashion, as exemplified in Fig. 1. Although they are definitely appealing alternatives, these frameworks are highly sensitive to *uncertain* pseudo labels. Moreover, there exists no study to jointly train the feature extraction networks and cost aggregation networks in a complementary and boosting manner.

2.2. Confidence-aware Contrastive Learning

In this section, we first study how to achieve better pseudo labels for dense correspondence, and then present a confidence-aware contrastive learning.

We start from classic uncertainty measurement based on forward-backward consistency checking as proposed in [20, 21, 22], where an argmax operator was applied twice for forward and backward directions, respectively. Specifically, the pseudo matching map $F^{\mathcal{C}}_{s\to t}$ from the matching cost $\mathcal{C}$, warping I_s toward I_t, is defined as follows:

$$F^{\mathcal{C}}_{s\to t}(i) = \text{argmax}_{j'}\, \mathcal{C}\,(i,j') - i \tag{1}$$

where j' is defined for all the points in the target. Similarly, $F^{\mathcal{C}}_{t\to s}$ can be computed, warping I_t toward I_s. In an non-occlusion region, we get a backward flow vector $F^{\mathcal{C}}_{t\to s}$ in the inverse direction as the forward flow vector $F^{\mathcal{C}}_{s\to t}$. If this consistency constraint is not satisfied, the points in the target are occluded at the matches in the source, or the estimated flow vector is incorrect. These constraints can be defined such that

$$\begin{aligned}\|F^{\mathcal{C}}_{s\to t}(i) + F^{\mathcal{C}}_{t\to s}(i + F^{\mathcal{C}}_{s\to t}(i))\|^2 \\ < \alpha_1(\|F^{\mathcal{C}}_{s\to t}(i)\|^2 + \|F^{\mathcal{C}}_{t\to s}(i + F^{\mathcal{C}}_{s\to t}(i))\|^2) + \alpha_2.\end{aligned} \tag{2}$$

Since there may be some estimation errors in the flows, we grant a tolerance interval by setting hyper-parameters α_1 and α_2. A binary mask $M^{\mathcal{C}}$ is then obtained by such forward-backward consistency checking, representing a non-occluded region as 1 and an occluded region as 0.

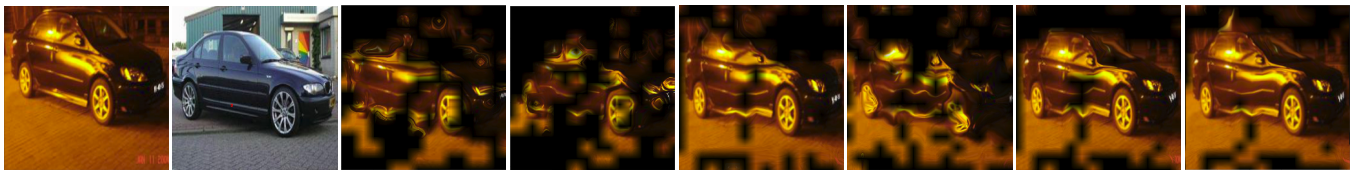
Based on the estimated mask $M^{\mathcal{C}}$, we present a confidence-aware contrastive loss function, aiming to maximize the similarities at the reliably matched points while minimizing for the others, defined such that

$$\mathcal{L}^{\mathcal{C}}_{\text{ccl}} = -\frac{1}{N^{\mathcal{C}}} \sum_i M^{\mathcal{C}}(i) \log \left(\frac{\exp(\mathcal{C}(i, i + F^{\mathcal{C}}_{s\to t}(i))/\gamma)}{\sum_j \exp(\mathcal{C}(i,j)/\gamma)} \right), \tag{3}$$

where $N^{\mathcal{C}}$ is the number of non-occluded pixels, and γ is a temperature hyper-parameter. Our loss function enables rejecting ambiguous matches with a thresholding while accepting the confident matches. While $\mathcal{L}^{\mathcal{C}}_{\text{ccl}}$ is formulated to train the feature extraction network itself, this loss function can also be defined for aggregated cost $\mathcal{A}$ as $\mathcal{L}^{\mathcal{A}}_{\text{ccl}}$, which can be used to train the cost aggregation networks.

2.3. Joint Learning

The proposed confidence-aware contrastive loss function can be independently used for learning feature extraction and cost



(a) I_s (b) I_t (c) $F_{s\to t}^{\mathcal{C},0}$ (d) $F_{s\to t}^{\mathcal{A},0}$ (e) $F_{s\to t}^{\mathcal{C},500}$ (f) $F_{s\to t}^{\mathcal{A},500}$ (g) $F_{s\to t}^{\mathcal{C},1000}$ (h) $F_{s\to t}^{\mathcal{A},1000}$

Fig. 2: **Comparison on PF-Pascal [18] applying masking and warping by the predicted correspondence map.** At later iteration (1000) with $F_{s\to t}^{\mathcal{A},1000}$ have more confident correspondences than with $F_{s\to t}^{\mathcal{C},1000}$ and at earlier iteration (500) vice versa.

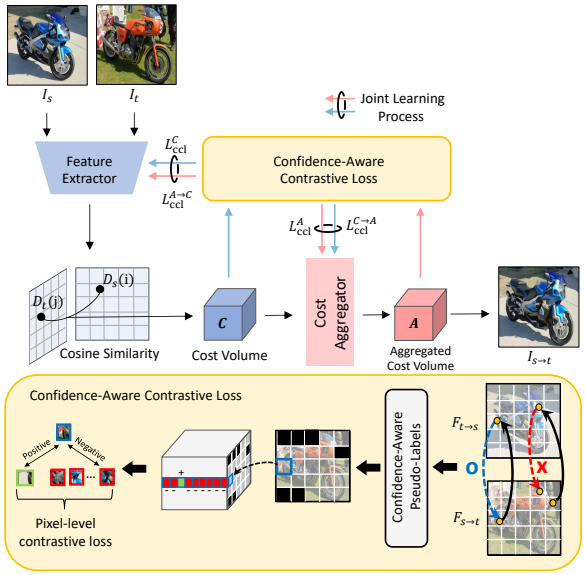


Fig. 3: **Overall framework.**

aggregation networks through $\mathcal{L}_{\text{ccl}}^{\mathcal{C}}$ and $\mathcal{L}_{\text{ccl}}^{\mathcal{A}}$, respectively. However, since two pseudo labels from feature extraction networks and cost aggregation networks may have complementary information, using two pseudo labels in a joint manner can enable further boosting the performance. For instance, at early stages of training, the pseudo label by *pre-trained* feature extractor provides more reliable cues than ones by *randomly-initialized* cost aggregation networks, which may help the cost aggregation networks converge much faster, as exemplified in Fig. 2. In addition, as the training progresses, the *well-trained* cost aggregation networks produce superior correspondences than the pseudo label by feature extractor, as exemplified in Fig. 2.

To leverage complementary information during training, we use a pseudo label output of each module, namely feature extraction and cost aggregation modules, defined such that

$$\mathcal{L}_{\text{ccl}}^{\mathcal{A}\to\mathcal{C}} = -\frac{1}{N^{\mathcal{A}}}\sum_{i} M^{\mathcal{A}}(i)\log\left(\frac{\exp(\mathcal{C}(i,i+F_{s\to t}^{\mathcal{A}}(i))/\gamma)}{\sum_j \exp(\mathcal{C}(i,j)/\gamma)}\right), \quad (4)$$

where $F_{s\to t}^{\mathcal{A}}(i) = \text{argmax}_{j'}\,\mathcal{A}(i,j')$. $\mathcal{L}_{\text{ccl}}^{\mathcal{C}\to\mathcal{A}}$ is similarly defined. Our final loss function is thus defined such that

$$\mathcal{L} = \lambda^{\mathcal{C}}(\mathcal{L}_{\text{ccl}}^{\mathcal{C}} + \mathcal{L}_{\text{ccl}}^{\mathcal{A}\to\mathcal{C}}) + \lambda^{\mathcal{A}}(\mathcal{L}_{\text{ccl}}^{\mathcal{A}} + \mathcal{L}_{\text{ccl}}^{\mathcal{C}\to\mathcal{A}}), \quad (5)$$

where $\lambda^{\mathcal{C}}$ and $\lambda^{\mathcal{A}}$ represent hyper-parameters. Fig. 3 illustrates the overall architecture of the proposed methods.

3. EXPERIMENTS

3.1. Implementation Details

In our framework, we used the ResNet-101 [25] pretrained on ImageNet [26] benchmark. We added additional layers followed by this to transform features to be highly discriminative w.r.t. both appearance and spatial context [4]. In addition, we used two types of cost aggregation modules like 4D CNN [5], denoted Ours w/NCNet, and transformer-based architecture [3], denoted Ours w/CATs. We used 256x256 size for the input image and 16x16 size for the feature map. The learning rate is adjusted, starting from differently 3e-5 and 3e-6 for feature extraction and cost aggregation, respectively, and adjusted using AdamW optimizer. We set $\alpha_1 = 0.1$, $\alpha_2 = 0.05$, $\lambda^{\mathcal{C}} = 0.5$, and $\lambda^{\mathcal{A}} = 0.5$.

3.2. Experimental Settings

In this section, we demonstrate that our method is effective through comparison with others; WeakAlign [10], RTNs [23], NCNet [5], DCCNet [24], ANCNet [13], PMNC [14], and CATs [3]. We also conduct an analysis of each component in our framework in the ablation study. To evaluate semantic matching, Proposal Flow [18] and TSS [19] benchmarks were used. The Proposal Flow benchmark contains PF-Pascal and PF-Willow [18]. TSS dataset is split into three subgroups: FG3D, JODS and PASCAL [19]. A percentage of correct keypoints (PCK) is employed for evaluation.

3.3. Experimental Results

Table 1 shows the quantitative results. We conduct experiments on various benchmarks, such as PF-Pascal [18], PF-Willow [18], and TSS [19]. Our method records higher accuracy than both baselines, NCNet and CATs, by 1.1 and 5.2 on PF-Pascal, 2.3 and 2.9 on PF-Willow, respectively. Specifically, ours w/CATs outperforms other methods over all

Method	Base	Trainable Comp.		Joint	PF-PSCAL(α = 0.1)	PF-Willow(α = 0.1)	TSS (α = 0.05)			
		Feature	Aggreg.				FG3D	JODS	PASCAL	Avg
WTA	ResNet-101	-	-	-	53.3	46.9	-	-	-	-
WeakAlign [10]	ResNet-101	✓	✓	-	75.8	84.3	90.3	76.4	56.5	74.4
RTNs [23]	ResNet-101	-	✓	-	75.9	71.9	90.1	78.2	63.3	72.2
NCNet [5]	ResNet-101	-	✓	-	78.9	84.3	94.5	81.4	57.1	77.7
DCCNet [24]	ResNet-101	✓	✓	-	82.3	73.8	93.5	**82.6**	57.6	77.9
ANCNet [13]	ResNet-101	✓	✓	-	86.1	-	-	-	-	-
PMNC [14]	ResNet-101	-	✓	-	90.6	-	-	-	-	-
CATs [3]	ResNet-101	✓	✓	-	87.3	76.9	85.3	73.7	55.4	73.6
Ours w/NCNet	ResNet-101	✓	✓	✓	80.0	**86.6**	**95.0**	82.3	55.8	78.4
Ours w/CATs	ResNet-101	✓	✓	✓	**92.5**	79.8	91.7	81.2	**60.9**	**80.0**

Table 1: **Comparison with state-of-the-art methods on standard benchmarks [18, 19].**

	Components	Accuracy
(a)	Ours	**80.0**
(b)	(-) Joint learning	78.4
(c)	(-) Confidence-aware loss	77.7

Table 2: **Ablation study on our modules.**

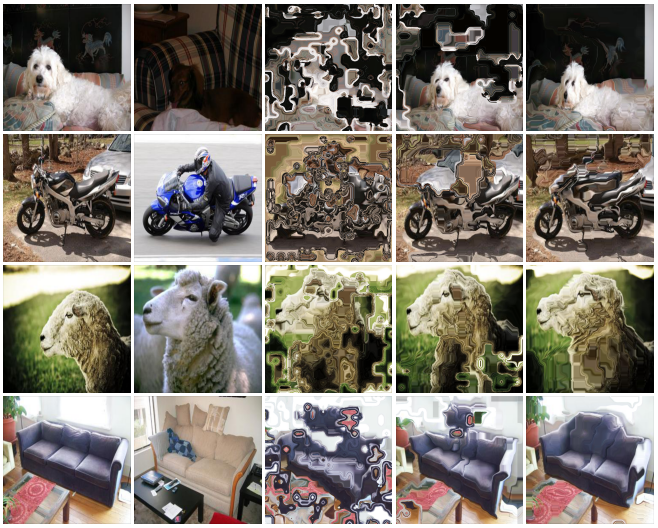
	Loss component				PCK (α = 0.1)
	$\mathcal{L}_{\text{ccl}}^{\mathcal{C}}$	$\mathcal{L}_{\text{ccl}}^{\mathcal{A}\rightarrow\mathcal{C}}$	$\mathcal{L}_{\text{ccl}}^{\mathcal{A}}$	$\mathcal{L}_{\text{ccl}}^{\mathcal{C}\rightarrow\mathcal{A}}$	
(a)	-	-	✓	-	70.0
(b)	-	-	✓	✓	71.7
(c)	✓	-	✓	-	78.4
(d)	✓	✓	✓	✓	**80.0**

Table 3: **Ablation study of our loss formulation.**

benchmarks and show the biggest performance improvement about 2 on PF-Pascal dataset compared to 90.6, which is the state-of-the-art result [14] among the similar network architectures and algorithms. Ours w/NCNet shows the highest performance on the FG3D as 95.0, and records average PCK of 78.4 on TSS benchmark. The qualitative results of semantic matching on the PF dataset are shown in Fig. 4. (c), (d), (e) are warped images from (a) to (b) by WTA, NCNet [5], and ours, respectively. Through (c), we could observe that errors of matches produced from feature extraction network affect the final output. Compared to (d), (e) shows accurate matching results even in difficult examples with occlusion, background clutter, and repetitive textures.

3.4. Ablation Study

In this section, we analyze the main components in our method, confidence-aware contrastive loss and joint learning, with NCNet baseline [5] on PF-Pascal. First, in Table 2 we validate the effectiveness of joint learning and confidence-aware contrastive loss by the lower performance of **(b)** and **(c)** compared to **(a)** which has both of these components. This proves that training two networks in a complementary manner boosts the performance and confidence-aware contrastive loss leads the unreliable pseudo labels to be filtered out during the training process. We also verify the effectiveness of each confidence-aware contrastive loss component through possible combinations of components displayed in Table 3.



(a) I_s (b) I_t (c) WTA (d) NCNet [5] (e) Ours

Fig. 4: **Qualitative results on PF-Pascal dataset [18].**

Compared to **(a)** and **(b)**, both of which train two networks only with the loss from the aggregation module, **(c)** and **(d)** show better PCK results by using separate losses that come from feature extraction and cost aggregation respectively. From this, we can confirm that the direct loss from each module works as a sufficient supervision signal for training which is free from gradient vanishing. Based on the performance improvements observed between **(b)** and **(a)**, and between **(d)** and **(c)**, we can also confirm that the reliable sample from one module helps training the other module, as it supports the formulation of better loss signals.

4. CONCLUSION

We address the limitations of the existing methods by jointly training feature extraction networks and aggregation networks in an end-to-end manner with the proposed confidence-aware contrastive loss. By jointly learning the networks with a novel loss function, our model outperforms the baseline and shows competitive results on standard benchmarks.

Acknowledgements. This research was supported by the National Research Foundation of Korea (NRF-2021R1C1C1006897).

5. REFERENCES

[1] Hajime Taira, Ignacio Rocco, Jiri Sedlar, Masatoshi Okutomi, Josef Sivic, Tomas Pajdla, Torsten Sattler, and Akihiko Torii, "Is this the right place? geometric-semantic pose verification for indoor visual localization," in *ICCV*, 2019.

[2] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely, "Learning feature descriptors using camera pose supervision," in *ECCV*, 2020.

[3] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim, "Semantic correspondence with transformers," *arXiv:2106.02520*, 2021.

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

[5] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic, "Neighbourhood consensus networks," *arXiv:1810.10510*, 2018.

[6] David G Lowe, "Distinctive image features from scale-invariant keypoints," in *IJCV*, 2004.

[7] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker, "Universal correspondence network," *arXiv:1606.03558*, 2016.

[8] 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.

[9] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic, "Convolutional neural network architecture for geometric matching," in *CVPR*, 2017.

[10] Ignacio Rocco, Relja Arandjelović, and Josef Sivic, "End-to-end weakly-supervised semantic alignment," in *CVPR*, 2018.

[11] Iaroslav Melekhov, Aleksei Tiulpin, Torsten Sattler, Marc Pollefeys, Esa Rahtu, and Juho Kannala, "Dgc-net: Dense geometric correspondence network," in *WACV*, 2019.

[12] Ignacio Rocco, Relja Arandjelović, and Josef Sivic, "Efficient neighbourhood consensus networks via submanifold sparse convolutions," in *ECCV*, 2020.

[13] Shuda Li, Kai Han, Theo W Costain, Henry Howard-Jenkins, and Victor Prisacariu, "Correspondence networks with adaptive neighbourhood consensus," in *CVPR*, 2020.

[14] Jae Yong Lee, Joseph DeGol, Victor Fragoso, and Sudipta N Sinha, "Patchmatch-based neighborhood consensus for semantic correspondence," in *CVPR*, 2021.

[15] Bastian Leibe, Aleš Leonardis, and Bernt Schiele, "Robust object detection with interleaved categorization and segmentation," in *IJCV*, 2008.

[16] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu, "Contrastive learning for unpaired image-to-image translation," in *ECCV*, 2020.

[17] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li, "Dense contrastive learning for self-supervised visual pre-training," *arXiv:2011.09157*, 2020.

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

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

[20] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer, "Dense point trajectories by gpu-accelerated large displacement optical flow," in *ECCV*, 2010.

[21] Simon Meister, Junhwa Hur, and Stefan Roth, "Unflow: Unsupervised learning of optical flow with a bidirectional census loss," in *AAAI*, 2018.

[22] Pengpeng Liu, Irwin King, Michael R Lyu, and Jia Xu, "Ddflow: Learning optical flow with unlabeled data distillation," in *AAAI*, 2019.

[23] Seungryong Kim, Stephen Lin, Sang Ryul Jeon, Dongbo Min, and Kwanghoon Sohn, "Recurrent transformer networks for semantic correspondence," in *NeurIPS*, 2018.

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

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[26] 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.