Cost Aggregation with 4D Convolutional Swin Transformer for Few-Shot Segmentation

Sunghwan Hong[1,⋆], Seokju Cho[1,⋆], Jisu Nam[1], Stephen Lin[2], and Seungryong Kim[1]

[1] Korea University, Seoul, Korea
{sung_hwan,seokju_cho,18wltnzzang,seungryong_kim}@korea.ac.kr
[2] Microsoft Research Asia, Beijing, China
stevelin@microsoft.com

In this document, we provide details on the experimental setting, more ablation studies, more quantitative results on semantic correspondence benchmarks, including SPair-71k [21], PF-PASCAL [6], and PF-WILLOW [5], and more qualitative results on all the benchmarks we used.

Appendix A. Experimental Setting for Semantic Correspondence

Datasets. For the datasets we used, we follow the common protocol [20, 22, 26, 36, 8, 19, 2] and use standard benchmarks [5, 6, 21]. Specifically, we consider SPair-71k [21], which provides a total of 70,958 image pairs with extreme and diverse viewpoints, scale variations, and rich annotations for each image pair. We also consider relatively small-scale datasets, which include PF-PASCAL [6] containing 1,351 image pairs from 20 categories and PF-WILLOW [5] containing 900 image pairs from 4 categories, where each dataset provides corresponding ground-truth annotations.

Evaluation metric. For evaluation on SPair-71k [21], PF-PASCAL [6], and PF-WILLOW [5], we employ the percentage of correct keypoints (PCK). It is computed as the ratio of estimated keypoints within the threshold from ground-truths to the total number of keypoints. Concretely, given predicted keypoint k_{pred} and ground-truth keypoint k_{GT}, we count the number of predicted keypoints that satisfy the following condition: $d(k_{\text{pred}}, k_{\text{GT}}) \leq \alpha \cdot \max(H, W)$, where $d(\cdot)$ denotes Euclidean distance; α denotes a threshold value; H and W denote height and width of the object bounding box or the entire image, respectively. We evaluate on PF-PASCAL with α_{img}, and SPair-71k, and PF-WILLOW with α_{bbox} following the common protocol.

Implementation Details. We use ResNet-101 [7] pre-trained on ImageNet [3] for the backbone feature extraction networks. We leave all the components in VAT unchanged. However, we build a different objective function. As in [20, 22, 19], we assume ground-truth keypoints are provided. We utilize Average End-Point Error (AEPE) [29] and compute it by averaging the Euclidean distance

⋆ Equal contribution

between the ground-truth and estimated flow. Specifically, we compute the loss as $\mathcal{L} = \|F_{\text{GT}} - F_{\text{pred}}\|_2$, where F_{GT} is the ground-truth flow field and F_{pred} is the predicted flow field. Note that we achieve this without making any modification to the network architecture. To report the results for different α thresholds, we employ the pre-trained weights released by authors, and simply evaluate without making any changes to their architectures. We use the same data augmentation used in CATs [2]. For the learning rate, we use the AdamW [17] optimizer with $3e^{-5}$ for VAT and $3e^{-6}$ for the backbone feature networks. Finally, we use appearance embedding from conv3_x, conv4_x and conv5_x as done for FSS-1000 [13].

Appendix B. Additional Ablation Study

Ablation study for feature backbone. Conventional few-shot segmentation methods only utilized CNN-based feature backbones [7] for extracting features. [34] observed that high-level features contain semantics of objects which could lead to overfitting and is not suitable to use for the task of few-shot segmentation. Then the question naturally arises, what about other networks? As addressed in many works [23, 4], CNN and transformers see images differently, which means that the kinds of backbone networks may affect the performance significantly, but this has never been explored for this task. We thus exploit several well-known vision transformer architectures to explore the potential differences that may exist.

Backbones feature	FSS-1000 [13] mIoU (%) 1-shot	5-shot
ResNet50 [7]	<u>90.1</u>	<u>90.7</u>
ResNet101 [7]	**90.3**	**90.8**
PVT [32]	90.0	90.6
Swin transformer [16]	89.8	90.2

Table 1. Ablation study of different feature backbone.

The results are summarized in Table 1. We find that both convolution- and transformer-based backbone networks attain similar performance. We conjecture that although it has been widely studied that convolutions and transformers see differently [23], as they are pre-trained on the same dataset [3], the representations learned by models are almost alike. Note that we only utilized backbones with a pyramidal structure, and the results may differ if other backbone networks are used, which we leave for future exploration.

Effectiveness of Data Augmentation. We explore the effectiveness of data augmentation for few-shot segmentation. In this experiment, we employ two types of data augmentation, which are introduced either in PFE-Net [28] or CATs [2]. We summarize the augmentation types in Table 4 and Table 3. For this ablation study, we use two datasets, PASCAL-5^i [27] and FSS-1000 [13]. The results are summarized in Table 2. Note that we use the same augmentation types and probability as theirs. For a fair comparison, we keep all the other experimental settings the same, *e.g.,* number of iterations and learning rate.

PFE-Net Aug. [28]	CATs Aug. [2]	PASCAL-5^i [27] mIoU (%)					FSS-1000 [13] mIoU (%)
		5^0	5^1	5^2	5^3	mean	
✗	✗	**70.0**	**72.5**	**64.8**	64.2	**67.9**	**90.3**
✓	✗	68.4	72.3	64.4	63.9	67.3	90.0
✗	✓	65.7	72.2	62.3	**64.3**	66.1	90.1
✓	✓	65.2	71.1	63.2	63.4	65.7	90.2

Table 2. Ablation study of Data Augmentation.

As PFE-Net [28] does not address the effectiveness of data augmentation and CATs [2] is designed for the semantic correspondence task, we are the first to analyze the effectiveness of data augmentation in the few-shot segmentation setting. Overall, we observe that using the data augmentation techniques severely affects the overall performance. Interestingly, although the augmentation technique introduced by CATs [2] showed a significant performance boost in the semantic correspondence task, it attains the lowest mIoU when evaluated on PASCAL-5^i [27] and the second lowest for FSS-1000 [13]. The severe performance drop in PASCAL-5^i [27] indicates a detrimental influence of using CATs [2] data augmentation. However, given the small difference to the best performance (0.3%) for FSS-1000 [13], the results may differ in a retrial. For PFE-Net [28] data augmentation, we observe results to be on par with the best reported results. However, at fold 0, there is a large gap between them, which indicates the detrimental effects of data augmentation on performance. Using both augmentations results in a large performance drop for PASCAL-5^i [27], arguably due to the detrimental effects of both augmentations, but for FSS-1000 [13], we observe only a small difference.

	Augmentation type	Probability
(I)	ToGray	0.2
(II)	Posterize	0.2
(III)	Equalize	0.2
(IV)	Sharpen	0.2
(V)	RandomBrightnessContrast	0.2
(VI)	Solarize	0.2
(VII)	ColorJitter	0.2

Table 3. CATs [2] Aug. Type.

	Strong Aug. type	Probability
(I)	RandScale	1
(II)	Crop	1
(III)	Gaussian Blur	0.5
(IV)	Horizontal Flip	0.5
(V)	Rotate	0.5

Table 4. PFE-Net [28] Aug. Type.

Consequently, we conjecture that the detrimental effects on PASCAL-5^i [27] and seemingly trivial effects on FSS-1000 [13] could be attributed to a few reasons: First, as shown in Table 2, since the difference between the results of the non-data augmentation approach and the PFE-Net [28] augmentation approach

is only 0.6% for PASCAL-5^i, this may be due to the implementation details. For the training, we followed HSNet [18] to force randomness for diverse episode combinations, which may have made such a gap. Second, although the data augmentation may help transformers by providing inductive bias and addressing the heavy need for data, for few-shot setting, where the objective is to predict labels of unseen classes, the results may be different to that of semantic correspondence. It was demonstrated [2] that for semantic correspondence, data augmentation indeed helps to boost the performance, but a different problem formulation for few-shot segmentation may result in detrimental effects. Third, since we act on correlation maps, applying data augmentation may significantly affect the matching distribution at each pixel. Unlike those works directly working on feature refinement [28, 35], where adopting data augmentation has a direct influence on feature maps, VAT aggregates the correlation maps computed between the features extracted from augmented images, which may result in different effects (performance drop) when the objective is to predict unseen classes. Lastly, combining both augmentations may increase the difficulty of learning, which in turn impacts accuracy.

Ablation study for ATD. In this ablation study, we show a quantitative comparison between the proposed ATD and a decoder without transformers [30, 31, 33, 16] to find out whether the model benefits from the use of transformers for further cost aggregation and filtering with the aid of the appearance embedding. For convenience, we call this Appearance-aware Decoder (AD). To implement this, we only exclude the transformers within ATD and leave all the other components and training settings unchanged, *e.g.,* network architecture, hyperparameters, learning rate and number of iterations. As shown in Table 5,

Components	FSS-1000 [13]					
	mIoU (%)		FB-IoU (%)		mBA (%)	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Convolutions	87.3	88.8	92.2	93.2	66.8	67.2
Transformers	90.3	90.8	94.0	94.4	68.0	68.6

Table 5. Ablation study for ATD.

we observe a large performance gap between AD and ATD, which demonstrates that using transformer allows for more effective aggregation, filtering and integration of correlation maps and appearance embedding. More specifically, we observe a 3% mIoU difference and find similar differences for FB-IoU and mBA. Without using transformers, where only convolutions are used, we observe that the results are equal to that of (**VI**) in the ablation study for VAT. This indicates that meaningful aggregation may not have occurred. It should be noted that we observe highly competitive results for mBA for both approaches, confirming a positive effect from the high-resolution spatial structure of the appearance embedding.

Ablation study for VCM. For this ablation study, we aim to further support our claims that the VCM (overlapping convolutions) compensates for the lack of inductive bias and alleviates the detrimental effects caused at window boundaries. To this end, we use Linear transformer [10], Fastformer [33] and Swin Transformer [16] to validate the effectiveness. Note that we already reported the results for the ones with VCM, but we additionally provide FB-IoU and mBA results. For the implementation of VEM, we refer the readers to Algorithm 1.

Components	FSS-1000 [13]					
	mIoU (%)		FB-IoU (%)		mBA (%)	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
VEM + Linear Transformer [10]	87.0	87.4	90.7	91.0	65.0	64.9
VEM + Fastformer [33]	87.1	87.6	90.9	91.2	65.3	65.2
VEM + Swin Transformer [16]	89.9	90.5	92.9	94.0	67.8	68.2
VCM + Linear Transformer [10]	87.7	88.3	92.3	92.2	66.5	66.7
VCM + Fastformer [33]	87.8	88.2	91.8	91.9	66.4	66.4
VCM + Swin Transformer [16]	90.3	90.8	94.0	94.4	68.0	68.6

Table 6. Ablation study for VCM.

As shown in Table 6, we find a similar pattern to the results for VCM. Swin Transformer attained the best results, while Linear Transformer [10] and Fastformer [33] show similar results. Interestingly, when VCM is replaced with VEM, the performance difference for Swin Transformer and the other two differ substantially. Specifically, for Swin Transformer, the mIoU is 89.9% when equipped with VEM, which is a 0.4% performance drop and is a relatively lower drop compared to those of Linear Transformer and Fastformer. This could be due to the relative position bias that Swin Transformer provides, which the other two transformers lack. Furthermore, we suspect that the lower mIoU results could be explained by one of the following factors: simplified self-attention computation, local smoothness property of a correlation map, and consideration of spatial structure.

Appendix C. Limitations

An apparent limitation is that since our approach acts on correlation maps, we need to explicitly compute the global correlation maps and store them. This is indeed memory expensive, and increases with the spatial resolution of the correlation maps. Although we utilize a coarse-to-fine architecture, this does not make the training feasible when resolutions are high. Specifically, given a spatial resolution of feature maps at size 128×128, the resultant size of correlation maps is at least 128^4, and counting the level dimensions as well as other pyramidal levels p, it is difficult to train with a sufficient batch size even with NVIDIA GeForce RTX-3090 GPUs. This might limit the accessibility of this approach. We also visualize failure cases in Fig. 1.

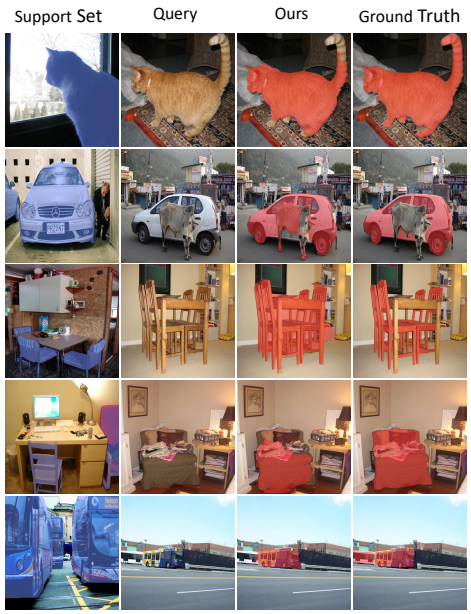

Fig. 1. Failure cases.

Appendix D. More Results

Quantitative Results for Semantic Correspondence. As shown in Table 7, we provide per-class quantitative results on SPair-71k [21] in comparison to other semantic correspondence methods, including CNNGeo [24], WeakAlign [25], NC-Net [26], HPF [20], SFNet [12], DCC-Net [8], GSF [9], SCOT [15], DHPF [22], CHM [19], MMNet [36], PMNC [11] and CATs [2].

Methods	aero.	bike	bird	boat	bott.	bus	car	cat	chai.	cow	dog	hors.	mbik.	pers.	plan.	shee.	trai.	tv	all
CNNGeo [24]	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
WeakAlign [25]	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
NC-Net [26]	17.9	12.2	32.1	11.7	29.0	19.9	16.1	39.2	9.9	23.9	18.8	15.7	17.4	15.9	14.8	9.6	24.2	31.1	20.1
HPF [20]	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 [15]	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 [22]	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 [19]	49.1	33.6	64.5	32.7	44.6	47.5	43.5	57.8	21.0	61.3	54.6	43.8	35.1	43.7	38.1	33.5	70.6	55.9	46.3
MMNet [36]	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	642	43.6	40.9
PMNC [11]	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
CATs [2]	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
VAT† (ours)	49.8	36.8	70.1	33.5	46.1	46.0	31.1	69.9	15.7	69.9	57.2	47.2	38.5	41.8	43.0	35.5	75.0	61.8	48.4
VAT (ours)	**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**

Table 7. Per-class quantitative evaluation on SPair-71k [21] benchmark.

More results for mBA comparison. In Table 8 and Table 9, we provide per fold quantitative results for mBA. Note that we obtained the mBA results for HSNet [18] and RePRI [1] using the pre-trained weights and code released by the authors. We omit the results for CyCTR [35] as the official code and weights by the authors are not publicly available.

Qualitative Results. As shown in Figure 2, Figure 3, Figure 4, Figure 5 and Figure 6, we provide qualitative results on all the benchmarks, which includes PASCAL-5^i [27], COCO-20^i [14], FSS-1000 [13], PF-PASCAL [6], PF-WILLOW [5] and SPair-71k [21].

Backbone network	Methods	1-shot					5-shot				
		5^0	5^1	5^2	5^3	mean	5^0	5^1	5^2	5^3	mean
ResNet50 [7]	RePRI [1]	45.8	53.7	46.6	50.0	49.0	45.4	46.9	41.8	41.0	43.8
	HSNet [18]	53.9	54.7	53.3	53.6	53.9	54.6	55.1	54.0	54.2	54.5
	VAT (ours)	55.1	55.1	53.8	53.6	54.4	55.4	55.3	54.5	53.9	54.8
ResNet101 [7]	RePRI [1]	47.6	47.6	41.9	43.3	45.1	46.4	44.4	38.4	38.7	42.0
	HSNet [18]	53.9	54.4	53.5	53.9	53.9	54.3	54.7	54.2	54.2	54.4
	VAT (ours)	54.7	54.6	53.9	55.5	54.7	55.0	55.0	54.5	54.8	54.8

Table 8. mBA comparison on PASCAL-5^i [27].

Backbone feature	Methods	1-shot					5-shot				
		20^0	20^1	20^2	20^3	mean	20^0	20^1	20^2	20^3	mean
ResNet50 [7]	RePRI [1]	6.84	6.16	5.76	6.46	6.31	5.44	4.45	3.49	3.47	4.21
	HSNet [18]	53.1	52.9	53.0	53.0	53.0	53.6	53.8	54.1	53.7	53.8
	VAT (ours)	54.1	54.0	54.5	54.0	54.2	54.6	54.8	55.4	54.7	54.9

Table 9. mBA comparison on COCO-20^i [14].

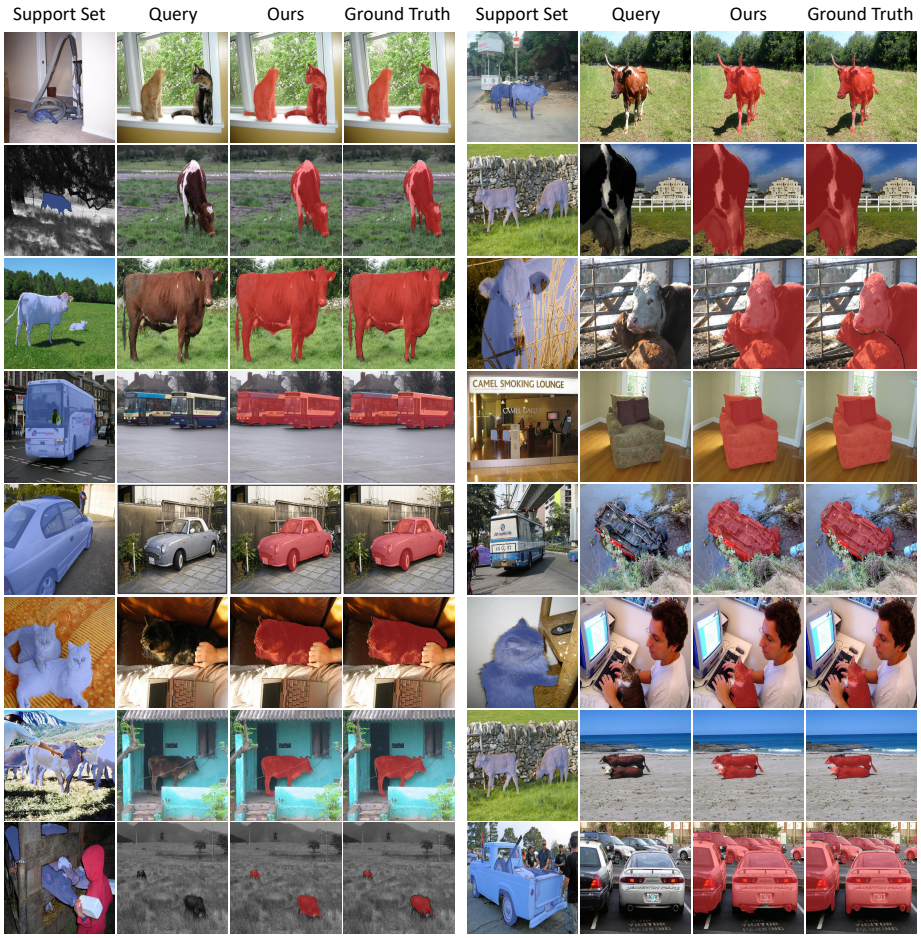

Fig. 2. Qualitative results on PASCAL-5^i [27].

Fig. 3. Qualitative results on COCO-20^i [14].

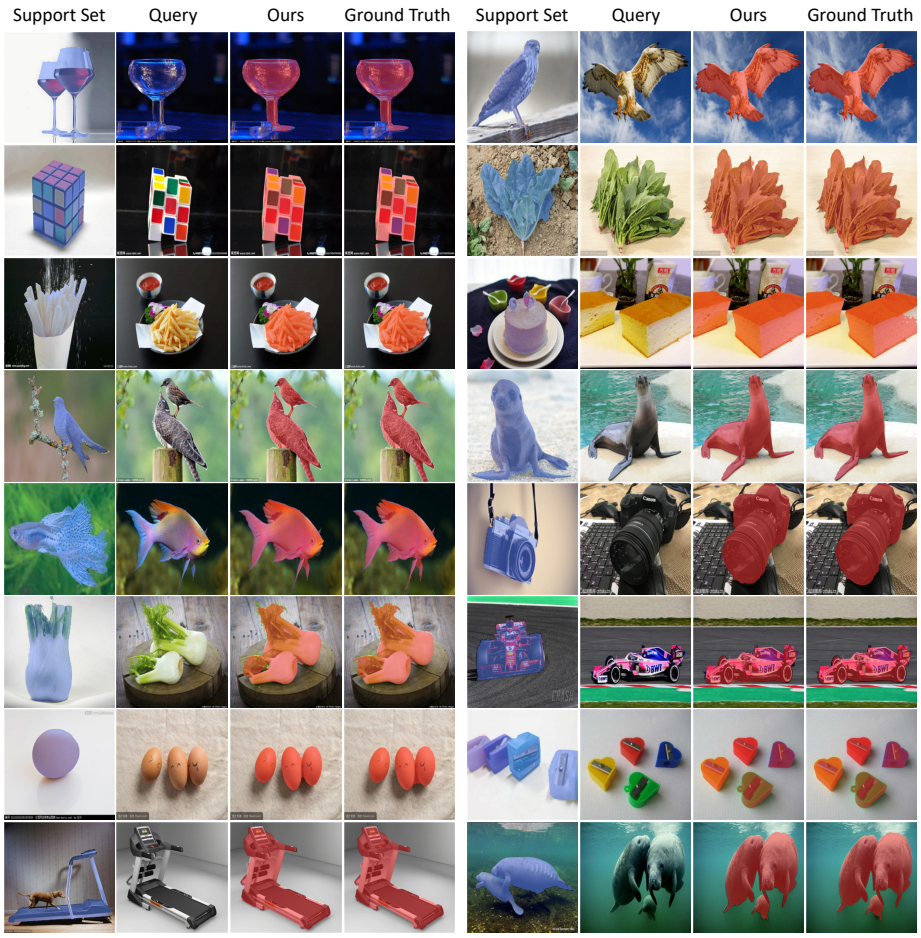

Fig. 4. Qualitative results on FSS-1000 [13].

Fig. 5. Qualitative results on PF-PASCAL [6] (left) and PF-WILLOW [5] (right).

Fig. 6. Qualitative results on SPair-71k [21].

References

1. Boudiaf, M., Kervadec, H., Masud, Z.I., Piantanida, P., Ben Ayed, I., Dolz, J.: Few-shot segmentation without meta-learning: A good transductive inference is all you need? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
2. Cho, S., Hong, S., Jeon, S., Lee, Y., Sohn, K., Kim, S.: Cats: Cost aggregation transformers for visual correspondence. In: Thirty-Fifth Conference on Neural Information Processing Systems (2021)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee (2009)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Ham, B., Cho, M., Schmid, C., Ponce, J.: Proposal flow. In: CVPR (2016)
6. Ham, B., Cho, M., Schmid, C., Ponce, J.: Proposal flow: Semantic correspondences from object proposals. IEEE transactions on pattern analysis and machine intelligence (2017)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)
8. Huang, S., Wang, Q., Zhang, S., Yan, S., He, X.: Dynamic context correspondence network for semantic alignment. In: ICCV (2019)
9. Jeon, S., Min, D., Kim, S., Choe, J., Sohn, K.: Guided semantic flow. In: ECCV. Springer (2020)
10. Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are rnns: Fast autoregressive transformers with linear attention. In: International Conference on Machine Learning. pp. 5156–5165. PMLR (2020)
11. Lee, J.Y., DeGol, J., Fragoso, V., Sinha, S.N.: Patchmatch-based neighborhood consensus for semantic correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
12. Lee, J., Kim, D., Ponce, J., Ham, B.: Sfnet: Learning object-aware semantic correspondence. In: CVPR (2019)
13. Li, X., Wei, T., Chen, Y.P., Tai, Y.W., Tang, C.K.: Fss-1000: A 1000-class dataset for few-shot segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
14. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision (2014)
15. Liu, Y., Zhu, L., Yamada, M., Yang, Y.: Semantic correspondence as an optimal transport problem. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

18. Min, J., Kang, D., Cho, M.: Hypercorrelation squeeze for few-shot segmentation. arXiv preprint arXiv:2104.01538 (2021)
19. Min, J., Kim, S., Cho, M.: Convolutional hough matching networks for robust and efficient visual correspondence. arXiv preprint arXiv:2109.05221 (2021)
20. Min, J., Lee, J., Ponce, J., Cho, M.: Hyperpixel flow: Semantic correspondence with multi-layer neural features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
21. Min, J., Lee, J., Ponce, J., Cho, M.: Spair-71k: A large-scale benchmark for semantic correspondence. arXiv preprint arXiv:1908.10543 (2019)
22. Min, J., Lee, J., Ponce, J., Cho, M.: Learning to compose hypercolumns for visual correspondence. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. Springer (2020)
23. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? arXiv preprint arXiv:2108.08810 (2021)
24. Rocco, I., Arandjelovic, R., Sivic, J.: Convolutional neural network architecture for geometric matching. In: CVPR (2017)
25. Rocco, I., Arandjelović, R., Sivic, J.: End-to-end weakly-supervised semantic alignment. In: CVPR (2018)
26. Rocco, I., Cimpoi, M., Arandjelović, R., Torii, A., Pajdla, T., Sivic, J.: Neighbourhood consensus networks. arXiv preprint arXiv:1810.10510 (2018)
27. Shaban, A., Bansal, S., Liu, Z., Essa, I., Boots, B.: One-shot learning for semantic segmentation. arXiv preprint arXiv:1709.03410 (2017)
28. Tian, Z., Zhao, H., Shu, M., Yang, Z., Li, R., Jia, J.: Prior guided feature enrichment network for few-shot segmentation. IEEE Transactions on Pattern Analysis & Machine Intelligence (2020)
29. Truong, P., Danelljan, M., Timofte, R.: Glu-net: Global-local universal network for dense flow and correspondences. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6258–6268 (2020)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems (2017)
31. Wang, S., Li, B.Z., Khabsa, M., Fang, H., Ma, H.: Linformer: Self-attention with linear complexity. arXiv preprint arXiv:2006.04768 (2020)
32. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. arXiv preprint arXiv:2102.12122 (2021)
33. Wu, C., Wu, F., Qi, T., Huang, Y., Xie, X.: Fastformer: Additive attention can be all you need. arXiv preprint arXiv:2108.09084 (2021)
34. Zhang, C., Lin, G., Liu, F., Yao, R., Shen, C.: Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5217–5226 (2019)
35. Zhang, G., Kang, G., Wei, Y., Yang, Y.: Few-shot segmentation via cycle-consistent transformer. arXiv preprint arXiv:2106.02320 (2021)
36. Zhao, D., Song, Z., Ji, Z., Zhao, G., Ge, W., Yu, Y.: Multi-scale matching networks for semantic correspondence. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3354–3364 (2021)