TransforMatcher: Match-to-Match Attention for Semantic Correspondence
— *Supplementary Material* —

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

In this supplementary material, we provide additional details, results and analyses of our proposed TransforMatcher pipeline.

A. Rotary positional embedding details

To keep the paper self-contained, we briefly explain on the formulation of rotary positional embedding (RoPE) [11]. The aim of RoPE is to find an encoding mechanism $f_{\{q,k\}}$ such that the inner product, g, of query q_m and key k_n of embeddings $\mathbf{x}_m, \mathbf{x}_n \in \mathbb{R}^d$ encodes position information only in the relative form as follows:

$$\langle f_q(\mathbf{x}_m, m), f_k(\mathbf{x}_n, n)\rangle = g(\mathbf{x}_m, \mathbf{x}_n, m-n), \quad (1)$$

where $m-n$ denotes the relative position between the embeddings. Starting from a simple case with dimension $d=2$, RoPE exploits the geometric properties of vectors on 2D plane and its complex form to prove that a solution to Eq. (1) is:

$$f_q(x_m, m) = (\mathbf{W}_q\mathbf{x}_m)e^{im\theta}, \quad (2)$$

$$f_k(x_n, n) = (\mathbf{W}_k\mathbf{x}_n)e^{in\theta}, \quad (3)$$

$$g(x_m, x_n, m-n) = \mathrm{Re}[(\mathbf{W}_q\mathbf{x}_m)(\mathbf{W}_k\mathbf{x}_n)^* e^{i(m-n)\theta}], \quad (4)$$

where $\mathrm{Re}[\cdot]$ is the real part of a complex number, $(\mathbf{W}_k\mathbf{x}_n)^*$ is the conjugate complex number of $(\mathbf{W}_k\mathbf{x}_n)$, and $\theta \in \mathbb{R}$ is a predefined non-zero constant. Writing $f_{\{q,k\}}$ in the form of matrix multiplication gives:

$$f_{\{q,k\}}(\mathbf{x}_m, m) = \begin{pmatrix} \cos m\theta & -\sin m\theta \\ \sin m\theta & \cos m\theta \end{pmatrix} \begin{pmatrix} \mathbf{W}^{(11)}_{\{q,k\}} & \mathbf{W}^{(12)}_{\{q,k\}} \\ \mathbf{W}^{(12)}_{\{q,k\}} & \mathbf{W}^{(22)}_{\{q,k\}} \end{pmatrix} \begin{pmatrix} x^{(1)}_m \\ x^{(2)}_m \end{pmatrix}, \quad (5)$$

where $[x^{(1)}_m, x^{(2)}_m]^\top = \mathbf{x}_m$ given $d=2$. Henceforth, to incorporate relative positional embedding, we can simply rotate the key/query embedding by amount of angle in multiples of its position index. The above formulation can be generalized to any even dimension d, by dividing the d-dimension space to $\frac{d}{2}$ sub-spaces, which are combined using the linearity of inner product. We refer the readers to the original paper [11] for full details.

B. Additional results and analyses

Category-wise PCK results. We show the category-wise PCK results of our model on the SPair-71k dataset [8] in comparison to existing methods in Table A1. It can be seen that TransforMatcher achieves the highest PCK overall, and the highest PCK in the majority of categories. An interesting observation is that while CATs [1] trained with augmentation shows consistently improved results compared to using no augmentation, TransforMatcher trained without augmentation often shows higher PCK values compared to TransforMatcher trained with augmentation. We conjecture this is because CATs also processes the actual 2D feature maps of source and target images together with the 4D correlation map using transformers, while TransforMatcher relies only on the 4D correlation map to find correspondences. An important takeaway is that is that while leveraging data augmentation provides more accurate semantic correspondences overall, it may have adverse effects on certain categories depending on the network architecture.

Ablation on correlation map channel dimension. We stated in the main paper that we construct a multi-channel correlation map as it is architecturally natural, and to exploit the richer semantics in different levels of feature maps. We conduct an experiment to compare the results of TransforMatcher when using a single-channel correlation map instead of a multi-channel correlation map. For fairness, we use the same bottleneck layers of `conv4_x` and `conv5_x`, and construct a single-channel correlation map by either (1) concatenating the multi-layer features along the channel dimension prior to correlation computation($\text{Single}_{\text{concat}}$), or (2) taking the mean of the multi-channel correlation map($\text{Single}_{\text{mean}}$). Table A2 shows the results of this comparison, where using multi-channel correlation map yields significantly higher results compared using a single-channel

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	dog	horse	mbike	person	plant	sheep	train	tv	all
NC-Net [10]	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 [7]	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 [5]	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 [9]	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 [6]	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 [4]	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 [12]	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 [1]	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	42.4
CATs† [1]	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	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†	**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**

Table A1. **Classwise PCK on SPair-71k.** Higher PCK is better. All the results reported in the table uses pretrained ResNet-101 model as the feature extractor. † indicates the use of data augmentation during training. Numbers in bold indicate the best performance, followed by the underlined numbers. It can be seen that while TransforMatcher achieves the highest PCK overall, the usage of augmentation results in a decrease in PCK in certain categories.

Channel	SPair-71k @α_{bbox} 0.05 (F)	SPair-71k @α_{bbox} 0.1 (F)
$\text{Single}_{\text{concat}}$	20.9	41.7
$\text{Single}_{\text{mean}}$	24.1	45.1
Multi (ours)	**32.4**	**53.7**

Table A2. **Ablation on correlation map channel dimension.** $\text{Single}_{\text{concat}}$ and $\text{Single}_{\text{mean}}$ denote single-channel correlation maps obtained by (1) concatenating the multi-layer features along the channel dimension prior to correlation computation, or (2) taking the mean of the multi-channel correlation map, respectively. Using multi-channel correlation map yields the highest results.

correlation map yielded by either $\text{Single}_{\text{concat}}$ or $\text{Single}_{\text{mean}}$.

C. Additional qualitative results

In Fig. A1, we qualitatively compare TransforMatcher and CATs [1], where TransforMatcher is seen to establish more accurate correspondences. We also show additional example visualization results in Figures A2-A4, where the source image is TPS-transformed [3] to the target image using predicted correspondences, aligning common instances in each image pair. As seen in Figures A2 and A3, the proposed method, TransforMatcher, effectively aligns foreground instances in presence of large scale, viewpoint, and illumination differences.

D. Details on nonlocality analysis of match-to-match attention

In this section, we provide implementation details regarding the analysis on nonlocality of match-to-match attention which is presented in the final part of section 5.2 of the main paper. Recall that we define the measure of nonlocality of an MHSA at layer l as the average of interactions between attention scores and relative offsets:

$$\Phi^l = \frac{1}{Z} \sum_{h \in [N_h]} \sum_{(\mathbf{q},\mathbf{k}) \in \mathcal{X} \times \mathcal{X}} \mathbf{A}^{(h)}_{\mathbf{q},\mathbf{k}} \|\mathbf{q} - \mathbf{k}\|^2, \tag{6}$$

where Z is normalization constant and $\mathcal{X}$ is a set of spatial positions in $\mathbf{C}$. As we found that the *global* query-key interaction in Eq.(5) is inadequate to effectively quantify this metric, we build *pair-wise* query-key interaction: $\mathbf{A}^{(h)}_{\mathbf{q},\mathbf{k}} = \sigma(\hat{\mathbf{Q}}^{(h)}\mathbf{K}^{(h)\top}) \in \mathbb{R}^{T \times T}$ where $\hat{\mathbf{Q}}^{(h)}_i := \mathbf{Q}^{(h)}_i \sigma(\tau \mathbf{w}_q \mathbf{Q}^{(h)\top})$, $\mathbf{q}, \mathbf{k} \in \mathbb{R}^4$, and $T = HWHW$. The further the query attends ($\|\mathbf{q} - \mathbf{k}\|$), the larger the nonlocality (Φ^l).

To measure the nonlocality of a convolutional layer, following the work of Cordonnier *et al.* [2], we represent a d-dim conv layer with kernel size K as an MHSA with K^d heads with following constraint: $\sigma(\mathbf{A}^{(h)}_{\mathbf{q},:})_{\mathbf{k}}$ equals to 1 if $\mathbf{q} - \mathbf{k} = \Delta_K$, and 0 otherwise where Δ_K is a set of local offsets. For example, $\Delta_K := [-1, 0, 1] \times [-1, 0, 1]$ if $d = 2$ and $K = 3$. We used $d \in \{4, 6\}$ and $K \in \{3, 5, 7, 9, 11\}$ in our experiments to visualize Figure 6.

In plotting Figure 7 of the main paper, we utilize the difficulty levels of image pairs in the SPair-71k dataset. Each pair in SPair-71k has annotations describing the types (viewpoint & scale variations, truncation, and occlusion) and levels (easy, medium, and hard) of difficulty. For truncation and occlusiuon, a pair is easy if no instances are truncated/occluded, medium if only one instance is, and hard if both are.

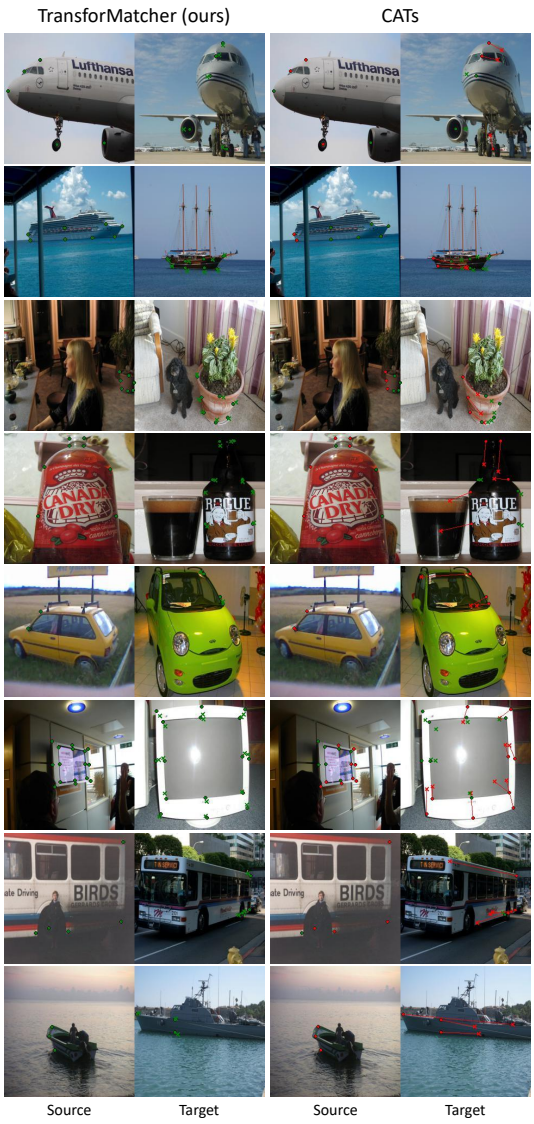

Figure A1. Qualitative comparison between the proposed TransforMatcher (left) and CATs [1] (right). We show keypoints in circles and predictions in crosses with a line that depicts matching error. Best viewed in electronic forms.

Figure A2. Example visualization results with large scale changes from SPair-71k [8].

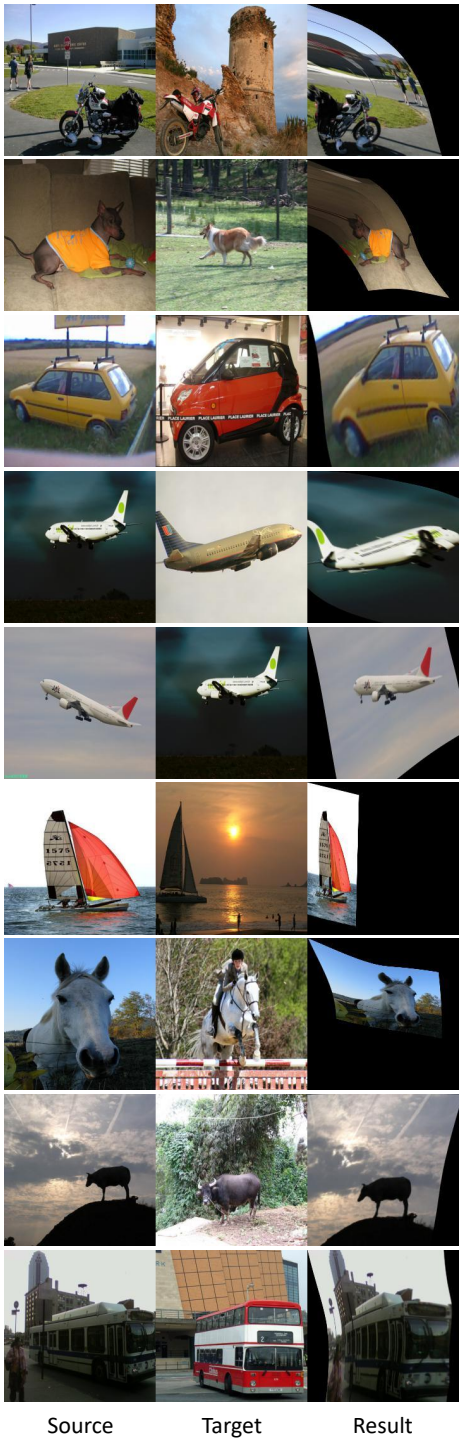

Figure A3. Example visualization results with large viewpoint and illumination changes from SPair-71k [8].

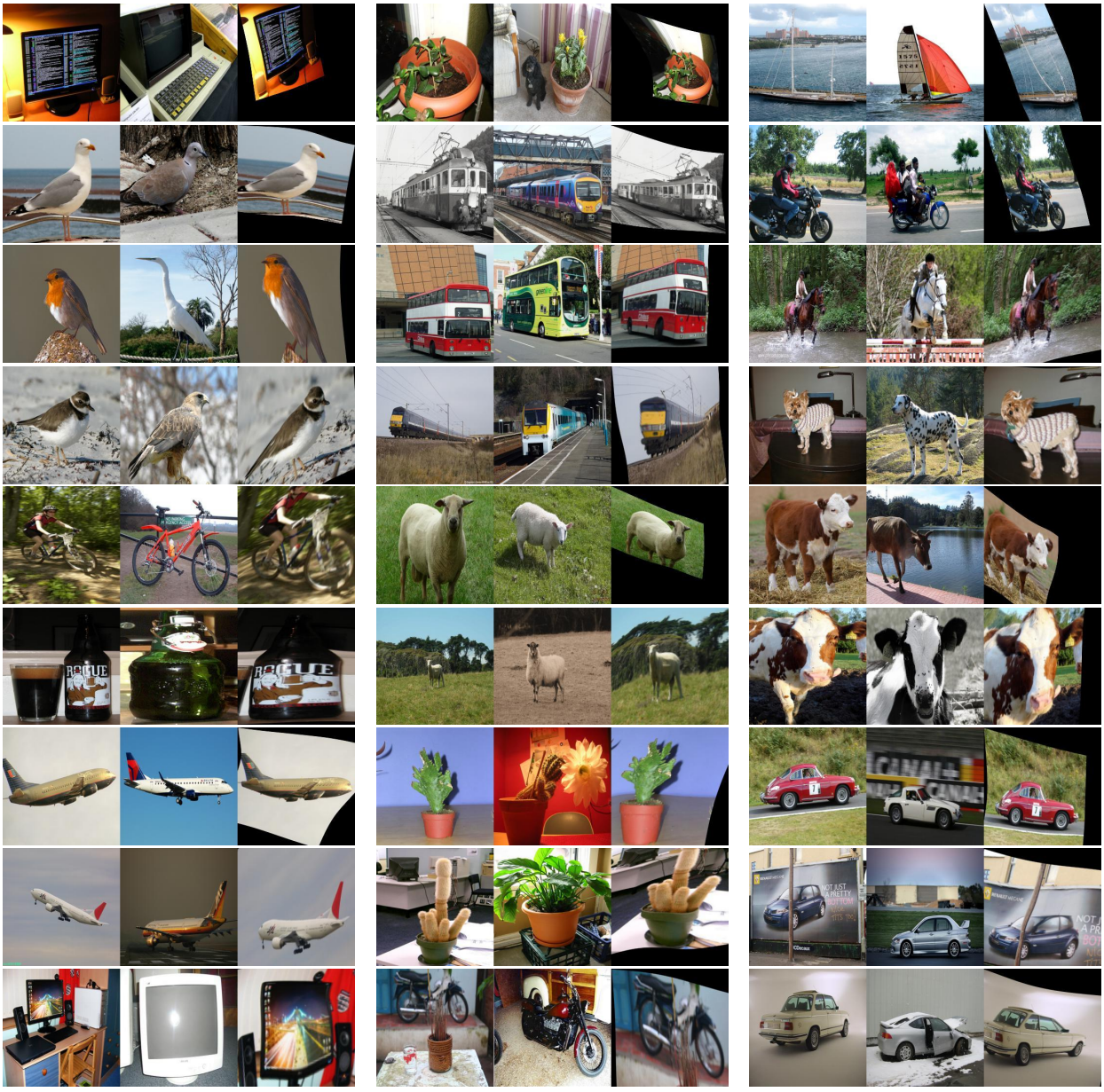

Figure A4. Example visualization results from SPair-71k [8].

References

[1] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Semantic correspondence with transformers. *arXiv preprint arXiv:2106.02520*, 2021. 1, 2, 3

[2] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolution. 01 2020. 2

[3] Gianluca Donato and Serge Belongie. Approximate thin plate spline mappings. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2002. 2

[4] Jae Yong Lee, Joseph DeGol, Victor Fragoso, and Sudipta Sinha. Patchmatch-based neighborhood consensus for semantic correspondence. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[5] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[6] 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. 2

[7] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 2

[8] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. SPair-71k: A large-scale benchmark for semantic correspondence. *arXiv prepreint arXiv:1908.10543*, 2019. 1, 3, 4, 5

[9] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Learning to compose hypercolumns for visual correspondence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[10] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[11] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. 1

[12] 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. 2