VISUAL TRANSFORMATCHER: EFFICIENT MATCH-TO-MATCH ATTENTION FOR VISUAL CORRESPONDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Establishing correspondences between images remains a challenging task, especially under large appearance changes due to different viewpoints and intra-class variations. In this work, we introduce a strong image matching learner, dubbed *Visual Transformatcher*, which builds on the success of the Transformers in vision domains. Unlike previous self-attention schemes over image matches, it performs match-to-match attention for precise match localization and dynamically updates matching scores in a global context. To handle a large number of candidate matches in a dense correlation map, we develop a light-weight architecture with an effective positional encoding technique for matching. In experiments, our method achieves the new state of the art on the SPair-71k dataset, while performing on par with existing state-of-the-art models on the PF-PASCAL and PF-WILLOW datasets, showing the effectiveness of the proposed approach. We also provide the results of extensive ablation studies to justify the design choices of our model. The code and trained weights will be released upon acceptance.

1 INTRODUCTION

Establishing correspondences between images is a fundamental task in computer vision, and is used for a wide range of problems including 3D reconstruction, visual localization and object recognition (Forsyth & Ponce, 2011). With the recent advances of deep neural networks, many learning-based keypoint extractors and feature descriptors were introduced (DeTone et al., 2018; Tian et al., 2019), showing significantly improved performances over their traditional counterparts. While further research addressed joint feature detectors and descriptors for sparse feature matching (Revaud et al., 2019; Dusmanu et al., 2019), dense feature matching methods have shown impressive performances despite higher computation complexities (Rocco et al., 2018). However, establishing reliable correspondences between images remains a challenging problem, especially under strong appearance differences, *e.g.*, viewpoint and illumination changes. In particular, the presence of intra-class variations, *i.e.*, scenes depicting different instances of the same category, remains a critical challenge for visual correspondence (Min et al., 2020; Liu et al., 2020; Min & Cho, 2021).

The CNN-based methods (Rocco et al., 2018; 2020; Min & Cho, 2021) evidence that refining and utilizing the 4D correlation map from feature matches between image pairs is essential to establish robust and accurate image correspondences. However, these methods suffer from the inherent limitations of convolutional layers, *i.e.*, receptive fields limited to the kernel size, and can only enforce semi-local geometric constraints on the correlation map or carry out geometric voting in a local manner. While convolutional neural networks have been the de-facto standard for not only visual correspondence but also for various other vision-related tasks, transformers have recently shown competitive results in vision domain, with reduced reliance on convolution layers. For example, Dosovitskiy et al. (2021) attain excellent results compared to convolutional baselines with fewer training computational resources; Vaswani et al. (2021) improve it in terms of both memory and time with local self-attention. These pioneering work show that transformers are attractive alternatives to convolutional layers in vision models, attributing to relaxed reliance on inductive biases, ability to easily scale to attend to global contexts, and dynamic attention.

Inspired by these accomplishments, we propose a novel image matching pipeline, dubbed *Visual TransforMatcher*, to tackle the challenging task of visual correspondence under intra-class variations using transformer networks. Specifically, we introduce match-to-match attention, a novel

mechanism to process correlation maps computed from features of images to match. Match-to-match attention considers the global match-wise interactions in the 4D space of correlation maps, refining and filtering matches in a more robust way which encompasses semi-local constraints as well. For practicality given the high number of tokens in the correlation map, we employ Fastformers (Wu et al., 2021) with additive attention for linear complexity, where we model the match-wise 4D positions using rotary positional embedding (Su et al., 2021). After refining the noisy correlation map with a series of match-to-match attention layers, we construct a dense flow map to transfer keypoints for establishing category-level correspondences between images.

Our contributions can be summarized as follows:

- We propose the Visual TransforMatcher, a novel image matching pipeline built on transformers for global-aware correlation map refinement and dynamic attentive weights,
- To the best of our knowledge, we are the first to directly process such a high-dimensional (4D) input using a self-attention mechanism within feasible computational constraints,
- We extend rotary positional embedding used in language sequences to model the 4D match-wise positional embedding, and
- We demonstrate state-of-the-art or comparable performances on standard benchmarks of category-level matching - PF-PASCAL, PF-WILLOW and SPair-71k.

2 Related work

Transformers for vision. The success of self-attention and transformers in NLP (Vaswani et al., 2017) has propagated to the area of computer vision, effectively replacing the entire deep convolutional pathways only with self-attention layers; Dosovitskiy et al. (2021) directly apply a Transformer architecture on non-overlapping medium-sized image patches for image classification, exhibiting impressive results when pretrained on a large-scale dataset. Touvron et al. (2021) introduce several training strategies that allow ViT to also be effective using the smaller ImageNet-1K dataset without large-scale pretraining. Vaswani et al. (2021) propose a new family of self-attention named HaloNets to improve the speed, memory usage and accuracy. While there are many other variants of vision transformers, *e.g.*, Liu et al. (2021); Wang et al. (2021), it has been shown that transformer-based vision models benefit from global context, relaxed inductive biases and dynamic attention weighting. We show that such characteristics are well applicable to the task of semantic matching domain as well, demonstrating importance of global receptive fields on match-to-match analyses.

Efficient Transformers. Due to the quadratic complexity of conventional transformers (Vaswani et al., 2017), they are infeasible to model extremely long-range interactions. This motivates the use of efficient transformers with lower computational complexity for feasible computation overhead when handling long sequences. Kitaev et al. (2020) reduces the complexity down to log-linear using locality-sensitive hashing and reversible residual layers. Wang et al. (2020) approximates the self-attention mechanism using low-rank matrices for linear complexity. Instead of relying on sparsity or low-rankedness, Choromanski et al. (2020) propose positive orthagonal random features approach (FAVOR+) to achieve linear complexity as well. Recently, Wu et al. (2021) proposed the Fastformer architecture which uses additive attention techniques only with element-wise products. In this paper, we choose to employ Fastformer to model match-to-match attention for not only its scalable complexity, but also for its simplicity and efficacy in modeling long-range interactions.

Positional Embeddings in Transformers. Positional embeddings aim to embed the position of each token as part of its features. Absolute positional embedding aims to be position-specific, where the embedding values can be predefined or learned (Sukhbaatar et al., 2015; Vaswani et al., 2017). Relative positional embedding aims to exploit relative positions (Shaw et al., 2018), so that the modelled interactions between tokens are relation-aware. However, conventional relative positional embedding requires an explicit computation of the attention matrix, which is absent in linear-complexity transformers. Su et al. (2021) propose rotary positional embeddings which inject spiral-like position information via multiplication for language sequences. In this work, we extend rotary positional embedding for effective 4D match-wise position embedding to encode relative position priors in our match-to-match attention mechanism.

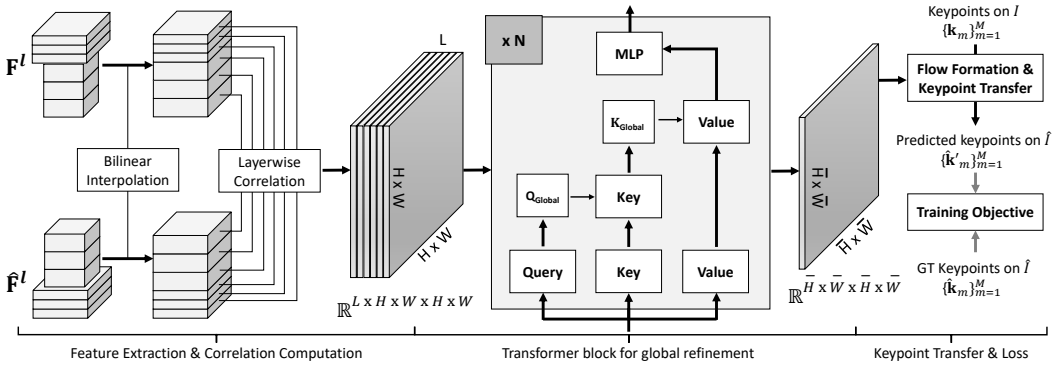

Figure 1: **Overview of Visual Transformatcher.** The feature maps extracted from an image pair are used to compute a multi-channel correlation map to be processed by our match-to-match attention module for refinement. We construct a dense flow field from the resulting correlation map, which can be used to transfer keypoints for training with keypoint pair annotation.

Category-level Matching. Category-level matching aims to find corresponding elements between images of different instances in the same category. Traditional approaches to category-level matching (Cho et al., 2015; Taniai et al., 2016) use hand-crafted descriptors to obtain matches between images. Recent approaches Min et al. (2020); Li et al. (2020); Jeon et al. (2020) build on the success of deep learning to extract learned features from convolutional neural networks, usually pretrained on the ImageNet classification task (Krizhevsky et al., 2012). An emerging trend is to exploit high-dimensional convolution on the correlation map, enforcing semi-local constraints on the correlation tensor to refine matches (Rocco et al., 2018; Lee et al., 2021a;b). However, these work commonly consume a high computational cost with a large number of parameters in the kernels, and only consider translation in space as an inherent property of convolutional kernels. Furthermore, Min & Cho (2021) extend the idea of probabilistic Hough matching (Cho et al., 2015) and propose an interpretable and light-weight high-dimensional kernel for visual correspondence to learn a reliable voting strategy instead of capturing diverse patterns. While these work have proven the efficacy of utilizing correlation maps to discover reliable matches between images, we propose that exploiting the global context of matches would show improved robustness and accuracy, especially under extreme appearance variation between the images to match. We therefore impose high-dimensional efficient attention on the 4D correlation map, exploiting the transformer architecture to easily scale to use the global context with dynamic attention weights for improved generalization.

A concurrent work, CATs (Cho et al., 2021), also employ the transformer network to model global consensus. However, they differ from our work in the following aspects: (1) We directly perform match-to-match attention on the 4D correlation map, but CATs perform two separate self-attention operations on reshaped correlation maps, (2) CATs additionally concatenates a linearly projected feature map with the correlation map, therefore using much more information, (3) Our method benefits from alleviated computational overhead per transformer layer, as we do not employ projected feature maps. This provides better potential for stacking more transformer layers in comparison.

3 Preliminaries: Transformer and Fastformer

Transformers are built on multi-head self-attention (MHSA) to model the contexts within a sequence by capturing the interactions between all pairs of inputs (Vaswani et al., 2017). MHSA consists of multiple self-attention layers, each of which takes the input tokens $\mathbf{X} \in \mathbb{R}^{T \times D_{\text{in}}}$ to form a global self-attention using linear projections of $\mathbf{W}_{\text{Q}}^{(h)}, \mathbf{W}_{\text{K}}^{(h)} \in \mathbb{R}^{D_{\text{in}} \times D_h}$ and $\mathbf{W}_{\text{V}}^{(h)} \in \mathbb{R}^{D_{\text{in}} \times D_v}$ to capture long-range dependencies between the tokens of the input sequence:

$$\text{Self-Attention}^{(h)}(\mathbf{X}) = \text{softmax}(\tau \mathbf{X}\mathbf{W}_{\text{Q}}^{(h)}(\mathbf{X}\mathbf{W}_{\text{K}}^{(h)})^{\top})\mathbf{X}\mathbf{W}_{\text{V}}^{(h)} \tag{1}$$

$$= \text{softmax}(\tau \mathbf{Q}^{(h)}\mathbf{K}^{(h)\top})\mathbf{V}^{(h)}, \tag{2}$$

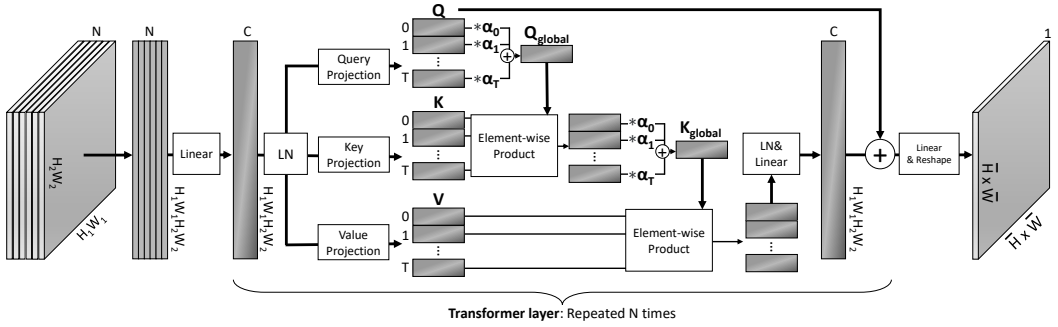

Figure 2: **Match-to-match attention module.** The multi-channel correlation map is projected to query, key and value matrices, which are multiplied with rotary positional embeddings. The match-to-match attention module exploits additive addition mechanisms to aggregate query/key matrices to global vectors, which is used for element-wise product to induce global context awareness. The final output is projected to a single-width channel to be reshaped to a refined 4D correlation map.

where (h) is the head index and τ is a scaling parameter. The MHSA with N_h heads aggregates the head outputs by an affine transformation layer with parameters $\mathbf{W}_O \in \mathbb{R}^{N_h D_v \times D_{\text{out}}}$ and $\mathbf{b}_O \in \mathbb{R}^{D_{\text{out}}}$:

$$\text{MHSA}(\mathbf{X}) = \underset{h \in [N_h]}{\text{concat}} \big[\text{Self-Attention}^{(h)}(\mathbf{X})\big] \mathbf{W}_{\text{O}} + \mathbf{b}_{\text{O}}, \tag{3}$$

It can be seen from this formula that the computational complexity of the transformer architecture is quadratic with respect to the sequence length T. This is a fundamental bottleneck of transformers when handling long sequences ($T > D_h$), which also pertains to our case of processing high-dimensional tensors, *i.e.*, pair-wise correlations between two 2-dimensional feature maps.

Fastformer aims to alleviate this bottleneck through the use of *additive* attention; instead of computing a quadratic attention map which encodes all possible interactions $\mathbf{Q}\mathbf{K}^\top \in \mathbb{R}^{T \times T}$, the fastformer (Wu et al., 2021) forms a compact key representation $\mathbf{P} \in \mathbb{R}^{T \times D_h}$ via additive attention which computes interactions between a global query representation and every key vector:

$$\mathbf{P}^{(h)}_{i,:} = \mathbf{K}^{(h)}_{i,:} \odot \sum_{j=1}^{T} \mathbf{Q}^{(h)}_{j,:} \text{softmax}(\tau \mathbf{w}_{\text{q}} \mathbf{Q}^{(h)\top})_j, \tag{4}$$

where $\mathbf{w}_{\text{q}} \in \mathbb{R}^{D_h}$ learns to transform the query vectors into a global vector. A similar additive attention mechanism summarizes the context-aware key representations $\mathbf{P}$ with a linear projection $\mathbf{w}_{\text{k}} \in \mathbb{R}^{D_h}$ to model its interaction with value vectors as follows:

$$\text{Self-Attention}^{(h)}_{\text{fast}}(\mathbf{X})_{i,:} = \mathbf{V}^{(h)}_{i,:} \odot \sum_{j=1}^{T} \mathbf{P}^{(h)}_{j,:} \text{softmax}(\tau \mathbf{w}_{\text{k}} \mathbf{P}^{(h)\top})_j, \tag{5}$$

with the assumption of $D_h = D_v$. The output is transformed by an MLP followed by residual connection with $\mathbf{Q}$. The empirical results (Wu et al., 2021) show that such additive attention can effectively model long-range interactions in the language domain while reducing the time and memory complexity down to linear: $\mathcal{O}(T^2 D_h) \rightarrow \mathcal{O}(T D_h)$.

4 VISUAL TRANSFORMATCHER

We provide an overview of our end-to-end matching pipeline. First, given a pair of images to match as an input, a feature extractor provides a set of intermediate feature pairs which are used to construct a multi-channel correlation map. This multi-channel dimension, compared to a single-channel dimension, ensures higher suitability to be projected to query, key and value matrices. Due to multifarious match-wise interactions within the global correlation map, we employ the linear-complexity Fastformers (Wu et al., 2021) to perform match-to-match attention, and extend the rotary positional embedding (Su et al., 2021) for match-wise 4D positional embedding. We refine the

multi-channel correlation map with several match-to-match attention layers, considering the global context within the correlation map for robust refinement. The refined correlation map is used to construct a dense flow field, which can be used for keypoint transfer to supervise our pipeline with ground-truth keypoint pair annotations. Fig. 1 illustrates the overview architecture of our method. The main components of our pipeline are detailed in subsequent subsections.

4.1 MULTI-CHANNEL CORRELATION COMPUTATION

We use the ImageNet-pretrained ResNet-101 architecture as a backbone feature extractor. We use all bottleneck layers of layers 3 and 4 to extract the features given an input pair of images $I, \hat{I} \in \mathbb{R}^{H \times W \times 3}$, and denote the set of intermediate feature pairs as $\{(\mathbf{F}^l, \hat{\mathbf{F}}^l)\}_{l=1}^{L}$. This is because in the case of transformers, it is architecturally natural for the input feature to have a sufficient dimension prior to being projected to key, query and value matrices with multiple heads.

The feature maps of the image pair to be matched, $\mathbf{F}^l, \hat{\mathbf{F}}^l \in \mathbb{R}^{H_l \times W_l \times D_l}$, are used to construct a correlation map $\mathbf{C}^l \in \mathbb{R}^{H_l \times W_l \times H_l \times W_l}$ if they are extracted from the same bottleneck layer, which represents the confidence score for all candidate correspondences between the two feature maps. Given a set of feature map pairs from different bottleneck layers $\{(\mathbf{F}^l, \hat{\mathbf{F}}^l)\}_{l=1}^{L}$, we compute the 4D correlation tensors for each pair as follows:

$$\mathbf{C}^l_{\mathbf{x},\hat{\mathbf{x}}} = \mathrm{ReLU}\Big(\frac{\mathbf{F}^l_{\mathbf{x},:} \cdot \hat{\mathbf{F}}^l_{\hat{\mathbf{x}},:}}{\|\mathbf{F}^l_{\mathbf{x},:}\| \|\hat{\mathbf{F}}^l_{\hat{\mathbf{x}},:}\|}\Big), \tag{6}$$

where $\mathbf{x}$,$\hat{\mathbf{x}} \in \mathbb{R}^2$ refer to 2-dimensional spatial positions of the feature maps corresponding to the image pair $(I, \hat{I})$. The L correlation tensors are then stacked together along the channel dimension after bilinear interpolation to the size of $H \times W \times H \times W$, *e.g.*, $\frac{1}{16}$ the size of the input image resolutions, resulting in the final correlation tensor $\mathbf{C} \in \mathbb{R}^{L \times H \times W \times H \times W}$.

This is unlike correlation maps used in prior work (Rocco et al., 2018), which only have a single channel, *i.e.*, one similarity score value for each pair of positions between the source and target feature maps. This is essential because having a single channel prior to the linear projection to query, key and value matrices would be ill-suited for a transformer-based architecture. Furthermore, leveraging different correlation tensors across the bottleneck layers allows us to exploit the richer semantics in different levels of feature maps, unlike previous methods which disregards the layer-wise similarities and semantics.

4.2 MATCH-TO-MATCH ATTENTION

We first flatten the 4D correlation map to behave as the sequence for the transformer module, *i.e.*, $\mathbb{R}^{L \times H \times W \times H \times W} \rightarrow \mathbb{R}^{L \times HWHW}$, considering similarity scores at each spatial position as a token embedding. The quadratic complexity of conventional self-attention in transformers poses an infeasible computation overhead in our setting, as a flattened 4D tensor results in a significantly long 1D tensor. We therefore use the recently proposed FastFormer (Wu et al., 2021), which proposes an *additive* self-attention mechanism with a linear computational complexity.

We first linearly embed the channel dimension of our flattened correlation map, *i.e.*, $\mathbf{X} = \mathbf{C}^\top \mathbf{W}_{\text{in}}$, where $\mathbf{C}$ refers to the correlation map obtained from the previous stage, $\mathbf{W}_{\text{in}} \in \mathbb{R}^{L \times D_{\text{in}}}$ is the linear transformation matrix, and $\mathbf{X} \in \mathbb{R}^{HWHW \times D_{\text{in}}}$ is the input to the Fastformer model. Instead of adding absolute positional encoding to $\mathbf{X}$ prior to the Multi-head self attention module, *e.g.*, $\mathbf{X}+E_{\text{pos}}$, we opt to integrate rotary positional embeddings (Su et al., 2021). The query, key and value are used for the additive self-attention mechanism, followed by an MLP and residual connection as in vanilla transformers. We use the pre-LN approach, where the layer normalization is placed inside the residual block for both the self-attention and MLP steps. Furthermore, we employ multi-head attention to ensure that our transformer module can attend to parts of the flattened correlation map differently. While the MHSA formulation is equal to Eq. (3), we formally formulate our fastformer-based self-attention module as follows:

$$\text{Self-Attention}^{(h)}_{\text{fast}}(\mathbf{C})_{i,:} = \sigma(\mathbf{X}\mathbf{W}^{(h)}_{\text{V}}; E_{\text{pos}})_{i,:} \odot \sum_{j=1}^{HWHW} \mathbf{P}^{(h)}_{j,:} \text{softmax}(\tau \mathbf{w}_{\text{k}} \mathbf{P}^{(h)\top})_j, \tag{7}$$

where $\mathbf{P}$ refers to the global context-aware key matrix obtained using Eq. (4) with $\mathbf{K} = \sigma(\mathbf{X}\mathbf{W}_{\mathrm{K}}; E_{\text{pos}})$ and $\mathbf{Q} = \sigma(\mathbf{X}\mathbf{W}_{\mathrm{Q}}; E_{\text{pos}})$, and $\sigma(\cdot; E_{\text{pos}})$ is a function that applies rotary positional embeddings to the projected input tokens. In a nutshell, our global consensus module takes as input a noisy correlation map to refine it using local and global context information to output a refined correlation map for robust image matching. A linear projection layer takes the concatenated outputs of the self-attention following Eq. (3). This process is repeated N times, providing a tensor in $\mathbb{R}^{L\times HWHW}$. The output from the transformer module is linearly projected to a single channel dimension, and is reshaped back to 4D: $\mathbb{R}^{L\times HWHW} \rightarrow \mathbb{R}^{H\times W\times H\times W}$, noise-filtered similarity scores for a reliable keypoint transfer. For precise transfer, we perform a 4-dimensional upsampling function on the 4D correlation map, and denote the tensor as $\mathbf{C}^{\text{out}} \in \mathbb{R}^{\bar{H}\times\bar{W}\times\bar{H}\times\bar{W}}$ where $\bar{H} = 2H$ and $\bar{W} = 2W$ which corresponds to $\frac{1}{8}$ the size of the original image. We illustrate the outline of our transformer-based global consensus module in Figure 2.

4.3 Flow field formation

The output correlation tensor $\mathbf{C}^{\text{out}}$ can be transformed into a dense flow field by applying kernel soft-argmax (Lee et al., 2019). We first normalize the raw correlation scores using softmax function:

$$\mathbf{C}^{\text{norm}} = \frac{\exp(\mathbf{G}^{\mathbf{p}}_{kl}\mathbf{C}^{\text{out}}_{ijkl})}{\sum_{(k',l')\in\bar{H}\times\bar{W}}\exp(\mathbf{G}^{\mathbf{p}}_{k'l'}\mathbf{C}^{\text{out}}_{ijk'l'})}, \tag{8}$$

where $\mathbf{G}^{\mathbf{p}} \in \mathbb{R}^{\bar{H}\times\bar{W}}$ is a 2-dimensional Gaussian kernel centered on $\mathbf{p} = \arg\max_{k,l}\mathbf{C}^{\text{out}}_{i,j,k,l}$, which is applied to smooth the potentially irregular correlation tensor values. The above equation returns a probability map $\mathbf{C}^{\text{norm}}$, which we use to transfer all the coordinates on the dense regular grid $\mathbf{P} \in \mathbb{R}^{\bar{H}\times\bar{W}\times 2}$ of source image I to obtain their corresponding coordinates $\hat{\mathbf{P}}' \in \mathbb{R}^{\bar{H}\times\bar{W}\times 2}$ on target image $\hat{I}$: $\hat{\mathbf{P}}'_{i,j} = \sum_{(k,l)\in\bar{H}\times\bar{W}}\mathbf{C}^{\text{norm}}_{i,j,k,l}\mathbf{P}_{k,l}$. We then can construct a dense flow field at sub-pixel level using the set of estimated matches $(\mathbf{P}, \hat{\mathbf{P}}')$.

4.4 Training objective

We assume that we are given the set of ground-truth coordinate pairs $\mathcal{M} = \{(\mathbf{k}_m, \hat{\mathbf{k}}_m)\}_{m=1}^{M}$ for each training image pair, where M is the number of annotated keypoint matches. We carry out keypoint transfer from the source keypoint to the target keypoint using the constructed dense flow field. A straightforward method of assigning a match $\hat{\mathbf{k}}'$ to some keypoint $\mathbf{k} = (x_k, y_k)$ would be to pick a single, discrete sample of a transferred coordinate *i.e.,* $\hat{\mathbf{k}}' = \hat{\mathbf{P}}'_{y_k x_k}$. However, this is likely to cause mislocalized keypoints as discrete sampling under sub-pixel level hinders fine-grained localization of keypoints. Therefore, for a given keypoint $\mathbf{k} = (x_k, y_k)$, we define a soft sampler $\mathbf{W}^{(k)} \in \mathbb{R}^{\bar{H}\times\bar{W}}$ as follows:

$$\mathbf{W}^{(k)}_{ij} = \frac{\max(0, \tau - \sqrt{(x_k - j)^2 + (y_k - i)^2})}{\sum_{i'j'}\max(0, \tau - \sqrt{(x_k - j')^2 + (y_k - i')^2})}, \tag{9}$$

where τ is a distance threshold, and $\sum_{ij}\mathbf{W}^{(k)}_{ij} = 1$. It can be seen that the soft sampler effectively samples each transferred keypoint $\hat{\mathbf{P}}'_{ij}$ by assigning weights inversely proportional to the distance to $\mathbf{k}$. Using this soft sampler, we assign a match to the keypoint $\mathbf{k}$ as $\hat{\mathbf{k}}' = \sum_{(i,j)\in\bar{H}\times\bar{W}}\hat{\mathbf{P}}'_{ij:}\mathbf{W}^{(k)}_{ij}$, being able to achieve up to sub-pixelwise accurate keypoint matches. By applying this keypoint transfer method on the source keypoints, we obtain the predicted keypoint pairs on image $\hat{I}$: $\{(\mathbf{k}_m, \hat{\mathbf{k}}'_m)\}_{m=1}^{M}$ by assigning a match $\hat{\mathbf{k}}'_m$ to each keypoint $\mathbf{k}_m$ 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:

$$\mathcal{L} = \frac{1}{M}\sum_{m=1}^{M}\|\hat{\mathbf{k}}_m - \hat{\mathbf{k}}'_m\|_2^2. \tag{10}$$

Table 1: **Performance on standard benchmarks of semantic matching.** Higher PCK is better. All results except for ours are from (Min & Cho, 2021). Numbers in bold indicate the best performance, followed by the underlined numbers.

Method	SPair-71k PCK @ α_{bbox}		PF-PASCAL PCK @ α_{img}		PF-WILLOW PCK @ α_{bbox}	
	0.1 (F)	0.1 (T)	0.05	0.1	0.05	0.1
NC-Net (Rocco et al., 2018)	20.1	26.4	54.3	78.9	-	-
DCC-Net (Huang et al., 2019)	-	26.7	55.6	82.3	-	-
DHPF (Min et al., 2020)	27.7	28.5	56.1	82.1	50.2	**80.2**
PMD (Li et al., 2021)	26.5	-	-	81.2	-	-
UCN (Choy et al., 2016)	-	17.7	-	75.1	-	-
HPF (Min et al., 2019a)	28.2	-	60.1	84.8	-	-
SCOT (Liu et al., 2020)	35.6	-	63.1	85.4	-	-
SCNet (Han et al., 2017)	-	-	36.2	72.2	38.6	70.4
DHPF (Min et al., 2020)	37.3	27.4	75.7	90.7	49.5	77.6
DHPF† (Min et al., 2020)	39.4	-	-	-	-	-
NC-Net* (Rocco et al., 2018)	-	-	-	81.9	-	-
DCC-Net* (Huang et al., 2019)	-	-	-	83.7	-	-
ANC-Net (Li et al., 2020)	-	28.7	-	86.1	-	-
PMD (Li et al., 2021)	37.4	-	-	90.7	-	-
CHMNet (Min & Cho, 2021)	46.3	<u>30.1</u>	<u>80.1</u>	91.6	**52.7**	<u>79.4</u>
PMNC (Lee et al., 2021a)	<u>50.4</u>	-	**82.4**	90.6	-	-
CATs (Cho et al., 2021)	43.5	-	-	-	-	-
CATs† (Cho et al., 2021)	49.9	27.1	75.4	**92.6**	50.3	79.2
TransforMatcher (ours)	50.2	**30.5**	78.9	90.4	<u>50.6</u>	76.3
TransforMatcher† (ours)	**53.1**	28.8	79.6	<u>92.4</u>	49.3	75.7

5 EXPERIMENTS

We evaluate our method on the task of category-level matching which aims to match semantically similar parts given an image pair of the same object category of different instances.

Datasets. We report our results on standard benchmark datasets of semantic correspondence: PF-PASCAL (Ham et al., 2018), PF-WILLOW (Ham et al., 2016) and Spair-71k (Min et al., 2019b). The PF-PASCAL and PF-WILLOW datasets are taken from four categories of the PASCAL VOC dataset, having small viewpoint and scale variations. The PF-PASCAL dataset contains 1,351 image pairs, which are augmented to produce 2,940 / 308 / 299 pairs for training, validation and testing, respectively. The PF-WILLOW dataset contains 900 image pairs, which are used for testing. The SPair-71k dataset contains 70,958 image pairs with diverse variations in viewpoint and scale, and is split to 53,340 / 5,384 / 12,234 image pairs for training, validation and testing, respectively. Not only is the SPair-71k dataset significantly larger in number, it also has more accurate and richer annotations regarding different levels of difficulty in occlusion, truncation, viewpoint and illumination.

Implementation details. Following recent methods (Min & Cho, 2021; Cho et al., 2021), we employ the ResNet-101 model pre-trained on the ImageNet dataset (Krizhevsky et al., 2012) as the feature extraction network. Note that the `conv4_x` and `conv5_x` layers in ResNet-101 have 23 and 3 bottleneck layers respectively, from which we extract feature maps to compute 26 layer-wise correlations maps for each image pair. We set the spatial size of the input image to 240×240, resulting in $H = W = 15$ for feature maps used for correlation computation, and $\bar{H} = \bar{W} = 30$. We use 6 transformer layers when training on the SPair-71k dataset, and 4 transformer layers when training on the PF-PASCAL dataset as the small dataset size of PF-PASCAL tends to overfit with high number of layers. Each of our match-to-match attention layers have 8 heads for multi-head self attention ($N_h = 4$), with head dimension of 4 ($D_h = D_v = 4$). After the correlation tensor is passed through the series of transformer layers for global context-aware refinement, it is finally passed through a fully-connected layer to output a correlation tensor with a single channel for dense flow formation, *i.e.*, $\mathbb{R}^{L \times HWHW} \rightarrow \mathbb{R}^{\bar{H} \times \bar{W} \times \bar{H} \times \bar{W}}$. The overall pipeline of our method is implemented using Py-

Torch (Paszke et al., 2019), and is optimized using the Adam optimizer with a constant learning rate of 1e-3. We finetune the feature extractor network at a lower learning rate of 1e-5.

Evaluation metric. We adopt the percentage of correct keypoints (PCK) for evaluation, which is the standard evaluation metric for category-level matching. Given a pair of ground-truth and predicted target keypoints $\mathcal{K} = \{(\hat{\mathbf{k}}_m, \hat{\mathbf{k}}'_m)\}_{m=1}^{M}$), PCK is measured by:

$$\text{PCK}(\mathcal{K}) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}[\|\hat{\mathbf{k}}_m - \hat{\mathbf{k}}'_m\| \leq \alpha_\tau \cdot \max(w_\tau, h_\tau)], \tag{11}$$

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

5.1 Results and analysis.

For the SPair-71k dataset, we evaluate two versions for our model: a finetuned model (F) which is trained on SPair-71k, and a transferred model (T) which is trained on PF-PASCAL. On the PF-PASCAL and PF-WILLOW datasets, we follow the common evaluation protocol to train our network on the training split of PF-PASCAL and evaluate on the test splits of PF-PASCAL and PF-WILLOW. We use the same training, validation, and test splits of PF-PASCAL used in Min & Cho (2021). The quantitative results are illustrated in Table 1.

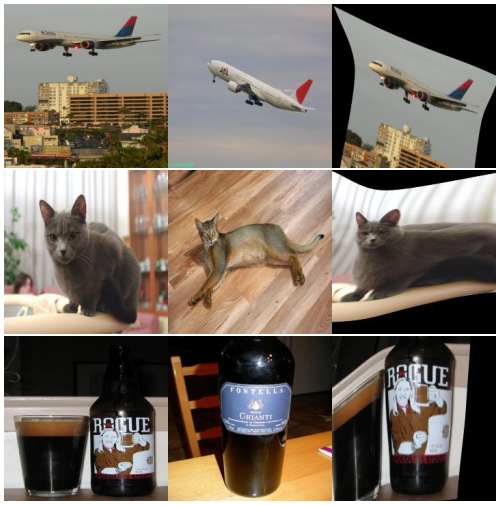

Figure 3: **Qualitative results on SPair-71k.** Source images are TPS transformed to target images using predicted correspondences.

We show that our proposed model finetuned on the SPair-71k dataset sets a new state of the art. A notable observation is that our model finetuned on the SPair-71k dataset without data augmentation outperforms CATs (Cho et al., 2021) trained with augmentation, proving the efficacy of our 4D match-to-match attention. Using data augmentations leads to improved PCK on both SPair-71k and PF-PASCAL datasets, but transformer-based models benefit more from augmentations as seen from the lower PCK increase in DHPF. It is interesting that our model trained without data augmentations transfer better to SPair-71k and PF-WILLOW datasets than our model trained with data augmentations, albeit its lower PCK performance on PF-PASCAL. This potentially hints that while data augmentations do help our transformer model to learn better, it overfits more to the training data domain, thus being less transferable to other data domains. Our model also exhibits state-of-the-art performance when transferred to the Spair-71k dataset, while being comparable on the PF-PASCAL and PF-WILLOW datasets. Figure 3 visualizes example qualitative results on SPair-71K using our model.

5.2 Ablation study and analysis

Correlation between Spair-71k and its *small* subset. Since the results on the PF-PASCAL dataset are nearly saturated, we use the SPair-71k dataset or its *small* subset for the experiments to justify our design choices. The *small* subset of SPair-71k contains 10,652 / 1,070 / 2,438 image pairs for training, validation, and test splits respectively. The results in Table 2 show that the trend of results using the *small* subset of SPair-71k is similarly consistent when using the large subset of SPair-71k.

Effect of data augmentation. Cho et al. (2021) found that using data augmentation for category-level matching model is beneficial, especially for data-hungry transformer-based architectures. We study the effect of applying data augmentation to our model as well. The results in Table 2 show that using data augmentation indeed gives consistent improvements to the performance of our model.

Table 2: **Ablation results on augmentation and positional embedding.** ✂ denotes *small* subset of SPair-71k. The results show that using data augmentation and rotary positional embedding gives the best results. Also, the trend of results from the *small* subset of SPair-71k is consistent in the standard SPair-71k dataset.

Augmentation	Pos. embedding	SPair-71k✂ PCK @ α_{bbox}		SPair-71k PCK @ α_{bbox}		PF-PASCAL PCK @ α_{img}	
		0.05	0.1	0.05	0.1	0.05	0.1
✗	Absolute (Ott et al., 2019)		-	74.5	89.4	29.9	48.7
✓	Absolute (Ott et al., 2019)	22.7	45.5	<u>79.4</u>	<u>91.9</u>	26.6	48.9
✗	Rotary (Su et al., 2021)		-	78.9	90.4	<u>30.5</u>	<u>50.2</u>
✓	Rotary (Su et al., 2021)	**30.1**	**51.8**	**79.6**	**92.4**	**32.3**	**53.1**

Table 3: **Ablation results on transformer architecture.** ✂ denotes *small* subset of SPair-71k. Vanilla transformers could not be evaluated within memory capabilities due to quadratic complexity. Fastformer shows the best results. Note that absolute positional embedding was used.

Transformer Architecture	SPair-71k✂ PCK @ α_{bbox}	
	0.05	0.1
Vanilla Transformer (Vaswani et al., 2017)	Out-Of-Memory	
Linformer (Wang et al., 2020)	0.5	1.6
Performer (Choromanski et al., 2020)	21.8	43.5
Fastformer (Wu et al., 2021)	**22.7**	**45.5**

Analysis on positional embedding. We investigate the effect of positional embedding used in our pipeline. As conventional relative positional embedding requires an explicit computation of the attention matrix, is not applicable to our transformer architecture with the additive attention. On the other hand, rotary positional embeddings can be seamlessly applied to our model as an alternative method to model relative positional embedding. The results in Table 2 show that using rotary positional embedding results in significant gains over absolute positional embedding.

Analysis on efficient transformer architecture. There exists other transformer architectures apart from Fastformers with various design choices to perform self-attention with (log)-linear complexity (Wang et al., 2020; Choromanski et al., 2020). We try replacing our Fastformer architecture with other efficient transformer designs, and also the vanilla transformer design to compare the performances. We use absolute positional embedding for its simple applicability. The results in Table 3 show that the Fastformer architecture shows the best results. We conjecture this is because other architectures rely on approximations or randomness to reduce the complexity of the original attention formulation, which could lead to inaccurate interactions between the position-sensitive similarity scores of the correlation maps. Experiments with Vanilla Transformers was infeasible due to its large memory demands of the pair-wise attention matrices.

6 Conclusion

In this paper, we have proposed the Visual TransforMatcher, an end-to-end transformer-based pipeline that performs category-level matching between images. Our principal contribution is our efficient match-to-match attention mechanism, which is, to the best of our knowledge, the first attempt to directly process a 4-dimensional input, *e.g.*, correlation map, using a transformer-based network with *global* receptive fields. This has been a challenging pursuit due to the quadratic complexity in modeling global-range interactions. Furthermore, we propose to extend the rotary positional embedding to blend with the 4D correlation map to provide high-dimensional position priors. The proposed model sets a new state of the art on the standard benchmarks of semantic matching, and we have performed various ablation tests to evidence our design choices of our approach. We anticipate this work will motivate the use of transformers with high-dimensional inputs in other domains.

REFERENCES

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. 2015.

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

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

Christopher Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. 2016.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 224–236, 2018.

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.

Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint detection and description of local features. *arXiv preprint arXiv:1905.03561*, 2019.

David Forsyth and Jean Ponce. *Computer Vision: A Modern Approach. (Second edition)*. Prentice Hall, November 2011. URL `https://hal.inria.fr/hal-01063327`.

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

Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. 2018.

Kai Han, Rafael S Rezende, Bumsub Ham, Kwan-Yee K Wong, Minsu Cho, Cordelia Schmid, and Jean Ponce. Scnet: Learning semantic correspondence. 2017.

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

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

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. 2012.

Jae Yong Lee, Joseph DeGol, Victor Fragoso, and Sudipta Sinha. Patchmatch-based neighborhood consensus for semantic correspondence, 2021a.

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)*, pp. 897–907, 2021b. doi: 10.1109/WACV48630.2021.00094.

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

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

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)*, pp. 7505–7514, June 2021.

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

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

Juhong Min and Minsu Cho. Convolutional hough matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2940–2950, June 2021.

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

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

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

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Jerome Revaud, Philippe Weinzaepfel, César De Souza, Noe Pion, Gabriela Csurka, Yohann Cabon, and Martin Humenberger. R2d2: Repeatable and reliable detector and descriptor. *arXiv preprint arXiv:1906.06195*, 2019.

Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. 2018.

Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Efficient neighbourhood consensus networks via submanifold sparse convolutions. 2020.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2074. URL https://aclanthology.org/N18-2074.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021.

Sainbayar Sukhbaatar, arthur szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/8fb21ee7a2207526da55a679f0332de2-Paper.pdf.

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

Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. Sosnet: Second order similarity regularization for local descriptor learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11016–11025, 2019.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers amp; distillation through attention. In *International Conference on Machine Learning*, volume 139, pp. 10347–10357, July 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12894–12904, June 2021.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. Fastformer: Additive attention can be all you need. *arXiv preprint arXiv:2108.09084*, 2021.