LocalTrans: A Multiscale Local Transformer Network for Cross-Resolution Homography Estimation

Ruizhi Shao, Gaochang Wu, Yuemei Zhou, Ying Fu, and Yebin Liu

1 Tsinghua University
2 Northeastern University
3 Beijing Institute of Technology

Abstract

Cross-resolution image alignment is a key problem in multiscale gigapixel photography, which requires to estimate homography matrix using images with large resolution gap. Existing deep homography methods concatenate the input images or features, neglecting the explicit formulation of correspondences between them, which leads to degraded accuracy in cross-resolution challenges. In this paper, we consider the cross-resolution homography estimation as a multimodal problem, and propose a local transformer network embedded within a multiscale structure to explicitly learn correspondences between the multimodal inputs, namely, input images with different resolutions. The proposed local transformer adopts a local attention map specifically for each position in the feature. By combining the local transformer with the multiscale structure, the network is able to capture long-short range correspondences efficiently and accurately. Experiments on both the MS-COCO dataset and the real-captured cross-resolution dataset show that the proposed network outperforms existing state-of-the-art feature-based and deep-learning-based homography estimation methods, and is able to accurately align images under $10 \times$ resolution gap.

1. Introduction

The rapidly development of multiscale gigapixel photography [5, 42, 45] brings large-scale, long-term and immersive visual experience. It synthesizes a single ultra-high-resolution image through aligning plenty of high-resolution local-views with a low-resolution global-view. In multiscale gigapixel photography, the large resolution gap between two views, namely cross-resolution, puts forward a new challenge to traditional homography estimation task. Homography estimation is defined as the estimation of the projection mapping between two views on the same plane in 3D space, which usually consists of three steps: feature
Deep learning-based homography estimation is first proposed by [8] using VGG-net [30] as backbone, which is more robust compared to traditional feature-based methods. To improve the generalization capacity on real data, Nguyen et al. [24] proposed an unsupervised learning method by minimizing a pixel-wise intensity error metric instead of the regression loss of homography matrix. To address the potential large motion

2. Related Work

In this section, we review topics on homography estimation, cross-resolution image alignment and attention mechanism that are the most relevant to our work.

Feature-based Homography Estimation. Methods in this category utilize feature points extracted from the image pair to obtain a set of feature correspondences. Then a homography is estimated based on the direct linear transform or the robust fitting algorithms such as RANSAC [9]. The accuracy of homography estimation depends on the quality of the detected image features. Traditional feature detectors, such as SIFT [18], SURF [2], ORB [29], are able to detect reasonable keypoints, which are robust to lighting, blur and perspective distortion. Recently, several deep learning-based feature extraction methods are also developed, e.g., LFNet [25] and ASLFeat [21], and reach a higher matching accuracy. Despite existing feature-based homography estimation methods can be robust to illuminance changes and foggy inputs, they often fail in cross-resolution cases.

Deep Homography Estimation. Deep learning-based homography estimation task.

Benefiting from the combination of the local transformer layer and the multiscale structure, the proposed LocalTrans network outperforms the state-of-the-art homography estimation methods in terms of PSNR and corner error on the MS-COCO dataset [17]. Moreover, we demonstrate that the LocalTrans network highlights a superior performance on challenging real-captured cross-resolution cases under resolution gap up to 10×, and further apply it to multiscale gigapixel photography, see Fig. 1. The main contributions are summarized as

- We propose to solve the cross-resolution problem in homography estimation using the transformer structure by explicitly capturing the correspondences between the inputs.
- We design a novel local transformer layer embedded within multiscale structure, which is able to capture correspondences with a long-short range attention. Experiments demonstrate that the proposed structure outperforms the global attention mechanism.
- The proposed local transformer has significantly faster speed and lower GPU memory cost compared with the vanilla transformer structure, achieving real-time homography estimation at 60fps (please see Table 1).
problem, Le et al. [16] proposed to use a cascade strategy to estimate the motion mask and the homography matrix in a coarse-to-fine manner. The above methods implicitly estimate the correlation between the two views by concatenating the images along the channel dimension. Alternatively, Zhang et al. [44] proposed to use a feature extractor with shared weights to extract image features separately, and directly concatenate the features in the following homography estimator network. However, this architecture is equivalent to concatenating the input images at the beginning of the network, and thus, fail to address the cross-resolution problem. Please refer to the comparison in Fig. 7 and Table 2.

**Cross-Resolution Image Alignment.** Cross-resolution image alignment is an open problem in cross-scale stereo [50], hybrid light field imaging [4, 48], multiscale gigapixel videography [42], etc. Commonly a pixel-to-pixel warping field is estimated and applied for registration between images with different resolutions. For example, Zhao et al. [47] presented a disparity estimation and refinement method for reconstructing high resolution light field in a hybrid light field imaging system. Zheng et al. [49] further proposed a deep learning-based optical flow estimation and image fusion method in a coarse-to-fine manner. To synthesize a multiscale gigapixel video, Yuan et al. [42] proposed an iterative feature matching and warping method to perform global and mesh-based homography estimations. In this paper, we investigate the cross-resolution problem in the homography estimation task by using transformer structure to pay attention to the correspondences across different resolution inputs.

**Attention Mechanism.** Attention was built to imitate the mechanism of human perception that mainly focuses on the salient part [13, 28, 7]. Vaswani et al. [33] indicated that the global attention mechanism is able to solve the long term dependency problem even without the backbone of a convolutional or a recurrent network. Wang et al. [36] introduced a self-attention to capture long-range dependencies (i.e., correspondences) by using matrix multiplication between reshaped feature maps. Alternative to the above global attention, many researches also investigated local attention approaches that focus on short-range dependencies (e.g., correspondences) by using matrix multiplication between reshaped feature maps. For example, Sperber et al. [31] introduces a soft Gaussian bias and a hard mask that is nonzero in a local region to control the context range attended by the network. Woo et al. [37] introduced a spatial attention module by using a convolution layer to extract inter-spatial relationship in a feature map.

Recently, numerous literatures show that the attention mechanism is efficient to obtain correlation across multimodal inputs, e.g., for visual question answering [3, 15, 41], video description [11] and texture transferring [39]. We therefore apply the attention mechanism to explicitly capture the correspondences across the input images, especially cross-resolution images, for homography estimation.
3. Methodology

3.1. Overview

In this paper, we introduce a novel deep homography estimation network with a multiscale local transformer structure, dubbed LocalTrans. The proposed LocalTrans network first applies a deep siamese network (i.e., an image encoder with shared weights) to extract features \( \phi_{IT} \) and \( \phi_{IU} \) from the target image \( I_T \) and the unaligned image \( I_U \), respectively. Two convolution layers followed by a \( 2 \times 2 \) max-pooling layer constitute a basic block in the deep siamese network. Therefore, we construct features with different scales in the multiscale structure by controlling the block number of the deep siamese network in each scale-level. In concrete details, we use \( K = k + 1 \) blocks to construct feature maps \( \phi^{(k)}_I \) of shape \( H_k \times W_k \) in scale level \( k \) (\( k = [1, 2, \ldots, K] \)), where \( H_k = \frac{H}{2^{k-1}}, W_k = \frac{W}{2^{k-1}} \), and \( H \) and \( W \) are the height and width of the input images.

Different from existing deep homography methods that simply concatenate the images or features, we then explicitly formulate correspondences between the features \( \phi_{IT} \) and \( \phi_{IU} \) by using the transformer structure (Sec. 3.2). In each scale-level, a homography estimation module (Sec. 3.3) is adopted to estimate a homography matrix \( H_k \) based on the attention map and feature maps. Then the unaligned image \( I_U \) is warped according to the homography matrix \( H_k \) and then fed into the next scale-level (please refer to Fig. 2 (a)).

3.2. Multiscale Local Transformer Network

In this section, we introduce the proposed local transformer incorporating a multiscale structure. A straightforward option to achieve a long-short range perception of correspondence between features is to implement the transformer structure [33] after the siamese network. However, the vanilla transformer structure brings high GPU memory and computational costs when processing high-dimensional features. To accelerate the transformer, we replace the global attention in the vanilla transformer structure by designing a novel local attention kernel (Sec. 3.2.2) and combining it with a multiscale structure. Despite the designed local transformer kernel only captures correspondences in a limited range in the low scale-level, the multiscale structure enables the network to perceive the correspondence in a long-short range manner.

3.2.1 Transformer Structure

The detailed architecture of the transformer structure in each scale-level is shown in Fig. 2(b). The inputs of the transformer are two features \( \phi^{(k)}_{IT} \) and \( \phi^{(k)}_{IU} \) output by the deep siamese network. Two modules, Self-Attention Encoder Module (SAEM) and Transformer Decoder Module (TDM), are employed to exploit internal relations within the feature maps via self-attention and to capture the correspondences across the two features from the multimodal inputs via the cross-attention, respectively.

**Self-Attention Encoder Module (SAEM).** The SAEM first applies three \( 1 \times 1 \) convolution layers \( f_Q(\phi_I), f_K(\phi_I) \) and \( f_V(\phi_I) \) without activation function to encode the input image feature \( \phi_I \) into features \( \phi_{Q,I}, \phi_{K,I}, \phi_{V,I} \) of shape \( C \times H_k \times W_k \), where \( C \) denotes the channel number. Then the self-attention result in the SAEM is computed as follows

\[
M_s = \sigma(\frac{\phi_{Q,I} \odot \phi_{K,I}}{\sqrt{C}}),
\]

\[
\phi_{h,I} = M_s \odot \phi_{V,I},
\]

where \( \sigma \) denotes the softmax function, and \( \odot \) and \( \odot \) denote the operations in the designed local transformer structure, which will be described in Sec. 3.2.2. The tensor \( M_s \) is usually interpreted as a self-attention map.

Since \( \phi_{Q,I}, \phi_{K,I}, \phi_{V,I} \) are derived from the same input \( \phi_I \), the self-attention mechanism encourages to enhance the edges and corners in the input feature \( \phi(I) \). As shown in an example in Fig. 3 (top), the network pays more attention to the edge with the same feature of the center pixel, which is more prominent in attention maps of higher scale-levels (e.g., \( M_s^{(2)} \) and \( M_s^{(3)} \)). The final high-level feature \( \phi_{s,I} \) output by the SAEM is generated by encoding the self-attention result \( \phi_{h,I} \) with a \( 1 \times 1 \) convolution layer.

**Transformer Decoder Module (TDM).** In this module, two iterations of cross-attention are used. In the first iteration, we first adopt three \( 1 \times 1 \) convolution layers \( f_Q(\cdot), f_K(\cdot), f_V(\cdot) \) without activation function to encode the high-level features \( \phi_{s,I} \) into features \( \phi_{Q,I}, \phi_{K,I}, \phi_{V,I} \) as those in the SAEM. But different with the self-attention in the SAEM (Eqn. 1), we apply a cross-attention mechanism between features from target image \( I_T \) and unaligned image \( I_U \), denoted as
\[ M_{I_U \rightarrow I_T} = \sigma\left( \frac{\phi^{T}_{Q,I_U} \odot \phi^{T}_{K,I_T}}{\sqrt{C}} \right), \]
\[ \phi^{h}_{h,I_T} = M_{I_U \rightarrow I_T} \odot \phi^{T}_{V,I_T}, \]
\[ M_{I_T \rightarrow I_U} = \sigma\left( \frac{\phi^{T}_{Q,I_T} \odot \phi^{T}_{K,I_U}}{\sqrt{C}} \right), \]
\[ \phi^{h}_{h,I_U} = M_{I_T \rightarrow I_U} \odot \phi^{T}_{V,I_U}, \]

where \( M_{I_U \rightarrow I_T} \) and \( M_{I_T \rightarrow I_U} \) are usually interpreted as cross-attention map, and \( \phi^{h}_{h,I_T} \) and \( \phi^{h}_{h,I_U} \) are attention-aware features.

In the second iteration, features \( \phi^{h}_{h,I_T} \) and \( \phi^{h}_{h,I_U} \) are first encoded into \( \phi^{h}_{s,I_T} \) and \( \phi^{h}_{s,I_U} \) using two 1 x 1 convolution layers, as shown in Fig. 2(b). Then we compute the attention map \( M_{att} \) using \( \phi^{h}_{s,I_T} \) and \( \phi^{h}_{s,I_U} \) as inputs

\[ M_{att} = \phi^{h}_{s,I_T} \odot \phi^{h}_{s,I_U}. \]

The attention map \( M_{att} \) serves to estimate the homography matrix, which will be introduced in Sec. 3.3.

The attention mechanism described in Eqn. 2 enables an interaction between target image \( I_T \) and unaligned image \( I_U \) in the feature space, and therefore, captures correspondence information more explicitly than simply feeding the image pair into the network through channel-wise concatenation [30, 8, 24, 16]. Besides, the network with the transformer structure is more robust when the input two images have considerably large resolution differences. In this cross-resolution case, more attention values (weights) are assigned to the same edge as that in the self-attention map \( M_{s,I_U} \).

### 3.2.2 Local Attention Kernel

To accelerate the transformer structure in Sec. 3.2, we propose a Local Attention Kernel (LAK) that captures correspondences in a local range, which is inspired by conventional 2D deconvolution (also known as transposed convolution) and convolution. The main difference is that the proposed local transformer applies variable slice in the attention map as the convolution kernel while the traditional 2D convolution or deconvolution adopts a fixed kernel in the feedforward path. In the following, we will introduce the proposed LAK by decomposing the attention mechanism in Eqn. 1 and 2 into two steps, as shown in Fig. 4.

**Local attention map generation.** Local attention map describes the correspondences between features \( \phi_Q \) and \( \phi_K \) in a local range, i.e., a squared window. Consider \( \phi_Q \in \mathbb{R}^{C \times H \times W} \), \( \phi_K \in \mathbb{R}^{C \times H \times W} \) and the radius of LAK is \( r \). Then for a certain element at position \( x = (x, y) \) in \( \phi_Q \) will first query the relationship with elements in a local range of \( N(x) \) with radius \( r \) in \( \phi_K \). Suppose the position of an element in \( \phi_K \) is \( u \in N(x) \), then a local correspondence map \( M' \) could be described as

\[ M'(x, u) = \phi^{T}_Q(x) \phi_K(u), \]

where \( \phi^{T}_Q(x) \in \mathbb{R}^{1 \times C} \) and \( \phi_K(u) \in \mathbb{R}^{C \times 1} \). The above formulation also explains the operation \( \phi_Q \odot \phi_K \) in Eqn. 1, Eqn. 2 and Eq. 3. And the final local attention map \( M \) is

\[ M = \sigma(\frac{M'}{\sqrt{C}}). \]

The above equation shows that the local attention map \( M \) is a 4D tensor with shape \( H_k \times W_k \times (2r+1) \times (2r+1) \), which records the correspondence of each position in \( \phi_Q \) with that in \( \phi_K \) in a local range. For instance, an element \( M(x, y, u, v) \) records the correspondence between point \( \phi_Q(x, y) \) and \( \phi_K(x+u, y+v) \), \( u, v \in [-r, r] \).

**Local attention convolution.** This operation uses the 4D local attention map \( M \) and the feature \( \phi_V \) to obtain high-level feature \( \phi_h \). Consider a 2D slice of \( M \) at a certain position \( x = (x, y) \), denoted as \( \omega_x \). The feature \( \phi_h \) is then obtained by performing the convolution between the feature \( \phi_V \) and the 2D slice \( \omega_x \)

\[ \phi_h(x) = \omega_x * \phi_V(x) = \sum_{u \in N(x)} M(x, u) \phi_V(x + u), \]

where \( M(x, u) \in \mathbb{R}^{1 \times 1} \) and \( \phi_V(x + u) \in \mathbb{R}^{C \times 1} \). The above formulation also explains the operation \( (\cdot) \odot \phi_V \) in Eqn. 1 and Eqn. 2.

**Discussions.** It should be noted that there is a clear difference between the proposed LAK and existing approach with local attention [10, 31, 38, 34, 37, 35]. Existing approaches typically implement local attention by setting an attention bias (a Gaussian bias [38] or a hard mask that is non-zero in a local region [31]) or convolution layers to perform channel squeezing [34, 37, 35]. In our proposed LoCalTrans, the convolution kernel in the LAK, i.e., the 2D slice \( \omega_x \), varies with the position \( x \), while the kernel in a conventional convolution is fixed in every position. We therefore term the operation as local attention convolution. The most related local attention structure was proposed in [22].
which computes a local weights within a small window and produce feature vector though weighted average. We generalize this concept to 2D space in combination with 2D convolution, making it more suitable for visual tasks.

Comparing with the global attention in the vanilla transformer structure, the proposed LAK effectively reduces the computational complexity from $O(H_K^2 \cdot W_K^2 \cdot C)$ to $O(H_k \cdot W_k \cdot (2r + 1)^2 \cdot C)$, and the memory usage of the attention map from $H_k^2 \cdot W_k^2$ to $H_k \cdot W_k \cdot (2r + 1)^2$. For a certain scale-level $k \in K$, we set the radius of the LAK $r = k + 1$ to encourage the local transformer to notice a longer range of correspondence in the higher levels.

To demonstrate the effectiveness of the proposed local attention, we compare the local transformer against the vanilla transformer with the same deep siamese network and the homography estimation module, as shown in Fig. 8. The result shows that the proposed LAK with even 1 scale is superior to the global attention in the vanilla transformer structure. In conclusion, the proposed local transformer kernel not only shows high computational efficiency but also has superior performance to the vanilla transformer structure with global perception of correspondences.

### 3.3. Homography Estimation Module

In each scale-level, the homography estimation module applies the attention map $\mathcal{M}_{att}$ in Eqn. 3 as input. A 8-dimensional vector is obtained by several convolution-pooling blocks to estimate the final homography matrix, as shown in Fig. 2(c).

More specifically, in a certain scale-level $k$, since the attention map $\mathcal{M}_{att}$ is a 4D tensor, we first reshape it from $[\frac{H}{2^k}, \frac{W}{2^k}, 2k + 3, 2k + 3]$ to $[(2k + 3)^2, \frac{H}{2^k}, \frac{W}{2^k}]$ (similar to the processing of cost volume in [32]) before feeding to the homography estimation module. Then we feed the feature into $k + 1$ convolution blocks, where each block contains two $3 \times 3$ convolution layers and one max-pooling layer. In the last block, the max-pooling layer is replaced by an average-pooling layer for generating a feature of shape $C$. For a 1D convolution layer (kernel size 1), the final output becomes a 8-dimensional vector, denoting the 2D offsets of 4 corner points $\hat{c}_i^{(k)}$, $\hat{c}_4^{(k)}$ in scale-level $k$. We can simply obtain the homography matrix $H_{k}$, from the offsets vectors and warp the the unaligned image $I_U^{(k)}$ using homography transform to obtain the input for the next level, i.e., $I_U^{(k+1)} = W a r p(I_U^{(k)}, H_{k})$.

Note that $H_{k}$ in each scale-level represents the homography matrix of the full resolution. Thus, the final homography matrix $H$ is computed directly by accumulating the estimated homography matrix $H_{k}$ in each level as follows

$$H = H_1 \times H_2 \times \cdots \times H_K.$$
Ablation study. To verify the efficiency of local transformer structure, we replace the proposed LAK with the vanilla (global) transformer in [33] while keeping the rest architecture of the network unchanged. This experiment is performed on an Intel(R) Xeon CPU E5-2699 V4 with 16GB memory and a NVIDIA RTX 2080 GPU. The comparison in Table 1 shows that the proposed LAK has higher computational efficiency in terms of both running time and GPU memory cost than the vanilla transformer. Moreover, results in Fig. 8 also validate the superior performance compared with a single scale global transformer network. Please refer to the supplementary for more ablation studies.

4.2. Synthesized Cross-Resolution Data

We compare our model on synthesized cross-resolution dataset with 6 baseline methods, a conventional feature-based method, SIFT+RANSAC [18], three deep learning-based methods, DHN [8], UDHN [44] and MHN [16], as well as two state-of-the-art Reference-based Super-Resolution (RefSR) methods, SRNTT [46] and TSR [39].

Fig. 7 shows the quantitative comparison on the MS-COCO dataset under 4× and 8× resolution gaps. The proposed LocalTrans network demonstrates superior performance on cross-resolution cases comparing with the conventional feature-based method, SIFT+RANSAC [18], and two deep learning-based methods, DHN [8] and MHN [16]. The numerical comparison in Table 2 shows that the proposed LocalTrans network significantly outperforms the
Figure 9. Visual evaluation on the multiscale gigapixel dataset [42] (top, 6×) and the cross-resolution stereo dataset [50] (bottom, 10×). We mix the GB channels of the aligned image and the R channel of the target image. The misaligned pixels appear as red or green ghosts.

| Method          | PSNR | SSIM | PSNR | SSIM |
|-----------------|------|------|------|------|
| SIFT+RANSAC [18]| 19.32| 0.802| 11.93| 0.64 |
| DHN [8]         | 19.64| 0.818| 17.04| 0.762|
| UDHN [44]       | 21.44| 0.864| 18.78| 0.834|
| MHN [16]        | 25.42| 0.951| 20.24| 0.860|
| SRNTT [46]      | 27.06| 0.961| -    | -    |
| TTSR [39]       | 27.89| 0.915| -    | -    |
| LocalTrans      | 30.17| 0.981| 24.12| 0.930|

Table 2. Numerical comparison (PSNR/SSIM) in different cross-scale settings on the MS-COCO datasets.

4.3. Optical Zoom-in Cross-Resolution Data

In this experiment, 4 baseline methods, SIFT+RANSAC [18], DHN [8], UDHN by Zhang et al. [44] and MHN [16], and the RefSR methods, SRNTT [46] and TTSR [39], are compared. On the multiscale gigapixel dataset [42], we then apply the local patches in a 5 × 5 grid of the global low-resolution image to estimate homography matrices separately. To ensure the spatial smoothness between neighbouring patches, we calculate the four corner points of each patch and take an average among its neighbors. The final result is obtained by warping each local high-resolution image to the corresponding grid as in [42].

Since there is no groundtruth for qualitative evaluation, we only demonstrate the visual comparison on the datasets [42, 50], as shown in Fig. 9. The resolution gaps between the local target images and the unaligned images are 6× in the multiscale gigapixel dataset [42] (top of Fig. 9) and 10× in the cross-resolution stereo dataset [50] (bottom of Fig. 9). The results show that the conventional SIFT+RANSAC [18] fails to estimate reasonable homography matrix in the first case, and the deep learning-based methods DHN [8], UDHN [44] and MHN [16] appear different degrees of missing alignments (please zoom-in for details). The proposed LocalTrans demonstrates the best visual results on the optical zoom-in cross-resolution datasets [42, 50], which has more complicated degradation model than synthesized cross-resolution data. More visual results on the optical zoom-in cross-resolution dataset are provided in the supplementary.

5. Conclusions

In this paper, we proposed a novel multiscale local transformer network, termed LocalTrans, for addressing the cross-resolution problem in homography estimation. We
consider the cross-resolution images as some kind of multi-modal input, and employ the transformer structure to explicitly capture correspondences between two modalities (cross-resolution images) in the feature space. To accelerate the transformer, we design a Local Attention Kernel (LAK) that generates a local attention map specifically for each position in the feature. By embedding the LAK within a multiscale structure, the proposed LocalTrans is able to capture correspondences in a long-short range manner. The proposed LocalTrans network outperforms state-of-the-art methods on the MS-COCO dataset and highlights a superior performance on the challenging real-captured cross-resolution dataset under resolution gap up to $10 \times$.

We believe our LocalTrans gives a new opportunity to learn robust and accurate interactions between cross-resolution inputs, and would be further applied to various applications, such as reference-based super-resolution, cross-resolution stereo matching and hybrid light field imaging.

References

[1] Daniel Barath, Jiri Matas, and Jana Noskova. Magsac: Marginalizing sample consensus. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10197–10205, 2019. 6

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In European conference on computer vision, pages 404–417. Springer, 2006. 2

[3] Hedi Ben-younes, Remi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 2631–2639, 2017. 3

[4] Vivek Boominathan, Kaushik Mitra, and Ashok Veeraraghavan. Improving resolution and depth-of-field of light field cameras using a hybrid imaging system. In IEEE International Conference on Computational Photography, 2014. 3

[5] David J. Brady, Michael E. Gehm, Ronald A. Stack, Daniel L. Marks, David S. Kittle, Dathon R. Golish, Esteban Vera, and Steven D. Feller. Multiscale gigapixel photography. Nature, 486(7403):386–389, 2012. 1

[6] Che-Han Chang, Chun-Nan Chou, and Edward Y Chang. Clkn: Cascaded lucas-kanade networks for image alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2213–2221, 2017. 6

[7] Maurizio Corbetta and Gordon L. Shulman. Control of goal-directed and stimulus-driven attention in the brain. Nature reviews neuroscience, 3(3):201–215, 2002. 3

[8] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. arXiv preprint arXiv:1606.03798, 2016. 2, 5, 6, 7, 8

[9] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6):381–395, 1981. 2

[10] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In Interspeech 2020, pages 5036–5040, 2020. 3, 5

[11] Chiiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R. Hershey, Tim K. Marks, and Kazuhiiko Sumi. Attention-based multimodal fusion for video description. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 4203–4212, 2017. 3

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015. 6

[13] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on pattern analysis and machine intelligence, 20(11):1254–1259, 1998. 3

[14] Yash Kant, Dhruv Batra, Peter Anderson, Alexander G. Schwing, Devi Parikh, Jiasen Lu, and Harsh Agrawal. Spatially aware multimodal transformers for text-vqa. In ECCV (9), pages 715–732, 2020. 2

[15] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In Advances in Neural Information Processing Systems, volume 31, pages 1564–1574, 2018. 3

[16] Huang Le, Feng Liu, Shu Zhang, and Aseem Agarwala. Deep homography estimation for dynamic scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7652–7661, 2020. 2, 3, 5, 6, 7, 8

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014. 2, 6, 7

[18] David G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2):91–110, 2004. 2, 6, 7, 8

[19] Zixin Luo, Tianwei Shen, Lei Zhou, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. Context前世: Local descriptor augmentation with cross-modality context. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2527–2536, 2019. 6

[20] Zixin Luo, Tianwei Shen, Lei Zhou, Siyu Zhu, Runze Zhang, Yao Yao, Tian Fang, and Long Quan. Geodesc: Learning local descriptor augmentation with cross-modality context. In Proceedings of the European Conference on Computer Vision (ECCV), pages 168–183, 2018. 6

[21] Zixin Luo, Lei Zhou, Xuyang Bai, Hongkai Chen, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. Asfleat: Learning local features of accurate shape and localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6589–6598, 2020. 2

[22] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1412–1421, 2015. 3, 5
[49] Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In Proceedings of the European Conference on Computer Vision (ECCV), pages 88–104, 2018. 3

[50] Yuemei Zhou, Gaochang Wu, Ying Fu, Kun Li, and Yebin Liu. Cross-mpi: Cross-scale stereo for image super-resolution using multiplane images. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 3, 6, 8