Perceptual Loss for Robust Unsupervised Homography Estimation

Daniel Koguciuk, Elahe Arani, Bahram Zonooz

Advanced Research Lab, NavInfo Europe, Eindhoven, The Netherlands

{daniel.koguciuk, elahe.arani}@navinfo.eu, bahram.zonooz@gmail.com

Abstract

Homography estimation is often an indispensable step in many computer vision tasks. The existing approaches, however, are not robust to illumination and/or larger viewpoint changes. In this paper, we propose bidirectional implicit Homography Estimation (biHomE) loss for unsupervised homography estimation. biHomE minimizes the distance in the feature space between the warped image from the source viewpoint and the corresponding image from the target viewpoint. Since we use a fixed pre-trained feature extractor and the only learnable component of our framework is the homography network, we effectively decouple the homography estimation from representation learning. We use an additional photometric distortion step in the synthetic COCO dataset generation to better represent the illumination variation of the real-world scenarios. We show that biHomE achieves state-of-the-art performance on synthetic COCO dataset, which is also comparable or better compared to supervised approaches. Furthermore, the empirical results demonstrate the robustness of our approach to illumination variation compared to existing methods.

1. Introduction

Given a pinhole camera model assumption, a homography relates any two images of the same planar surface in space or any two images produced by pure rotational movement of the camera [13]. One image, called a source, can be transformed by a 3×3 homography matrix H as if viewed from the viewpoint of the other image, called target. Even if the primary assumptions are violated, the homography can be applied as an initial alignment step in other tasks such as mesh flow [21] and optical flow [15]. Therefore, homography has been widely used in many computer vision applications such as SLAM [5, 26], image stitching [39, 12], and change detection and description [3, 40].

Traditionally, homography estimation was performed using non-learnable approaches either in the pixel-space (direct methods) or in hand-crafted feature space (feature-based methods) [39]. Recently, with the advancement in
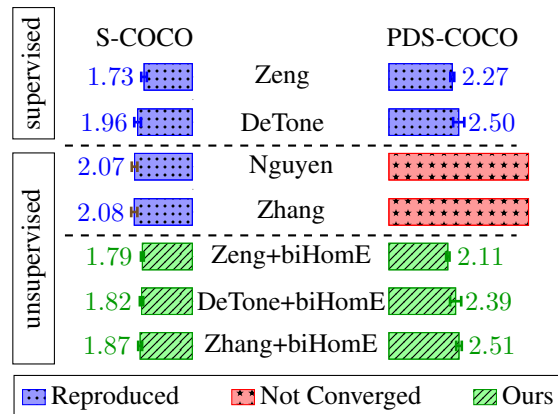

Figure 1. Homography estimation comparison on synthetic COCO (S-COCO) [4] on the left and a photometrically distorted version of it (PDS-COCO) on the right. On PSD-COCO, our bidirectional implicit Homography Estimation (*biHomE*) loss is the only unsupervised method able to converge, while still being on par with the performance of supervised approaches. The performance of methods (Nguyen [27], DeTone [4], Zeng [43], and Zhang [44]) is reported in Mean Absolute Corner Error (MACE).

Deep Neural Networks (DNNs), DeTone *et al.* [4] proposed a simple CNN architecture trained in an end-to-end fashion. The idea was to directly regress the parameters of a homography and it achieved similar performance to traditional feature-based methods. A more effective approach is presented by Zeng *et al.* [43], where the problem was formulated as per-pixel offset regression. However, supervised methods are often unlikely to be used in real-life scenarios, where ground truth homography labeling is prohibitively expensive.

To mitigate this issue, Nguyen *et al.* [27] introduced an end-to-end unsupervised approach. First, a homography is estimated using both input images and then homography estimation is learned by comparing the per-pixel intensity of warped source image and target image. Models trained using this formulation perform surprisingly well even for images with big viewpoint differences. However, they are not robust to big illumination changes and cannot be used in real-life scenarios [16, 29]. In contrast, Zhang *et al.* [44], by

learning the feature representation used for both homography estimation and image comparison, achieved robustness to different lighting conditions, but not for images with big viewpoint changes. Therefore, robust unsupervised homography estimation for both big illumination and viewpoint changes at the same time is still an open problem.

Instead of learning a shared representation for homography estimation and image comparison [44], we propose to decouple those two tasks by using a dedicated *Loss Network* to compare warped source image and target image. It has been shown that a pre-trained and fixed convolutional neural network as a *Loss Network* can capture perceptual differences [16, 41]. The homography is learned implicitly by comparing the images in feature space produced by the *Loss Network*. Since the homography should be invertible, we compare warped source and target images and swap their roles. Therefore, we call our loss function bidirectional implicit Homography Estimation (*biHomE*) loss.

We exhibit the simplicity and effectiveness of *biHomE* loss by applying it on top of three homography estimation architectures on the Synthetic COCO (S-COCO) dataset [4]. *biHomE* not only outperforms other unsupervised methods but is also on par with its supervised counterpart. Next, we use an additional photometric distortion augmentation on synthetic COCO (denoted as PDS-COCO) to mimic the illumination changes in real-life scenarios. We show that *biHomE* is the only unsupervised method that can converge on PDS-COCO, while still maintaining comparable or better performance compared to supervised approaches. To further demonstrate the effectiveness of our loss, we provide an experimental study on the influence of image alignment on the quality of generated change captions by the Dual Dynamic Attention Model (DUDA) [28] method. We achieve a new state-of-the-art performance on the change captioning task. Our contributions are as follows:

- We introduce a new perceptual loss (*biHomE*) to be used in an unsupervised setting, which decouples homography estimation from representation learning for image comparison.

- We propose to use an additional photometric distortion step in the synthetic COCO dataset generation (PDS-COCO) to evaluate the robustness of homography estimation to big illumination and viewpoint changes, which is more aligned with real-life scenarios.

- We achieve state-of-the-art performance using *biHomE* loss on both S-COCO and PDS-COCO datasets for unsupervised homography estimation.

- We apply *biHomE* for unsupervised image alignment to achieve a state-of-the-art change captioning quality on the CLEVR_Change dataset [28].

2. Related Work

2.1. Traditional Approaches

There are two families of traditional approaches for homography estimation: direct methods and feature-based methods [39]. Direct methods work in pixel intensity space, where most of the studies follow the Lucas-Kanade algorithm [24]. There are many improved approaches, including utilizing different error metrics [6] or Fourier domain representation [25]. Feature-based approaches mostly consist of three components: keypoint detection (such as SIFT [23] and ORB [34]), correspondence matching (Euclidean distance, correlation), and homography estimation using Direct Linear Transform (DLT) [13] with RANSAC [8] outlier rejection. Feature-based methods usually perform better than direct methods, however, their success depends on the quality of the hand-crafted stages and the content of the image itself. Poorly distributed features (*i.e.* repetitive pattern or texture), illumination variations, or large viewpoint differences [42] are challenging for both direct and feature-based methods.

2.2. Supervised Approaches

DeTone *et al*. [4] uses a deep learning model to estimate homography, which comes on par with or better than traditional methods. Input images are concatenated channel-wise and passed through 10-layer VGG-style [36] CNN with a fully connected layer with eight outputs (x-y coordinates for 4 points) on top. The problem is then formulated as a 4-point homography regression given the ground truth transformation. On the other hand, Zeng *et al*. [43] estimates pixel-to-pixel bijection between two images using a U-Net [33] architecture. The final homography during inference is produced using RANSAC and DLT in the postprocessing stage. Their method achieves a performance boost compared to DeTone's *et al*. [4] approach.

2.3. Unsupervised Approaches

Nguyen *et al*. [27] formulated the 4-point homography regression as an unsupervised approach. The idea is to minimize pixel-wise intensity error between the warped source image and the target image. They achieve comparable performance to the supervised DeTone *et al*. [4] method on the S-COCO dataset. Their method, however, cannot compensate for larger illumination changes between images as the learning is performed using pixel intensity comparison. Zhang *et al*. [44], applies the loss function to the feature space instead of the pixel-space. The authors propose to minimize the feature distance between the warped source and the target images and maximize the feature difference between the source and warped source images. Although presented triple loss is efficient for images with small viewpoint shifts, it is not robust to big viewpoint changes.

3. Method

There are two key insights of our method. The first one is that splitting the architecture into two separate components effectively decouples the homography estimation from representation learning. The second insight of our method is that we want to leverage the fact that convolutional neural networks pre-trained for image classification have already learned to encode meaningful information, which can be used in downstream tasks like image comparison for homography estimation assessment. Therefore, as shown in Figure 2, the system is composed of two components: The *Homography Estimation Network* (*HEN*) f, which can be realized by any architecture able to produce a transformation matrix H and a fixed *Loss Network* $g(I; \phi)$ that is used to define loss function L. Homography estimation is learned implicitly by minimizing the loss function defined in the feature space produced by *Loss Network* g.

3.1. Homography Estimation Networks

The goal of *HEN* is to estimate a 3×3 homography matrix between two images I_S and I_T using the learnable parameters θ:

$$H_{ST} = f(I_S, I_T; \theta) \tag{1}$$

Modern *HENs* can be divided into two categories: architectures that directly produce homography (DeTone *et al.* [4] and Zhang *et al.* [44]) and models that produce per pixel offset (Zeng *et al.* [43]). The former can be directly applied into our framework, but that is not the case with the latter. Thus, we first randomly sample without replacement $N_S < H \cdot W$ points out of the output perspective field of size $H \times W$ and then use Direct Linear Transform (DLT) [13] to estimate the homography matrix in an end-to-end pipeline.

3.2. Perceptual Loss Functions

In the second part of the pipeline, homography H_{ST} estimated by *HEN* is used to transform source image I_S into $I'_S = Warp(I_S, H_{ST})$. Rather than encouraging the pixels of the warped source image I'_S to match the pixels of the target image I_T (similar to [27]), we rely on perceptual similarity of high-level features extracted from pre-trained *Loss Network* g. Let $g_j(I)$ be the feature map of size $C_j \times H_j \times W_j$ after jth layer of the *Loss Network* g for a given image I. We define the mask M of the same size as $g_j(I)$ consisting of all ones and use $M' = Warp(M, H_{ST})$ to make sure the loss is applied only to the part visible on both images. The perceptual loss can be defined in multiple ways.

Implicit Homography Estimation with MAE/MSE Loss. The simplest approach is to define loss as per pixel normalized distance in feature space produced by g between
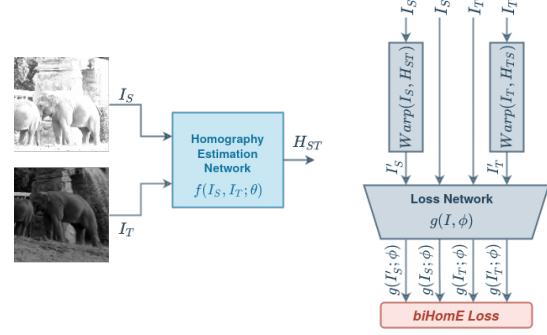

Figure 2. Proposed bidirectional implicit Homography Estimation (*biHomE*) loss, where f is any architecture producing homography H_{ST} between the input images I_S and I_T and g is a pretrained and frozen *Loss Network*. In our experiments, we use ResNet34 [14] as feature extractor g.

warped source image I'_S and target image I_T:

$$L_d(I_S, I_T) = \frac{\sum_{i}^{H_j, W_j, C_j} M'_i \cdot d(g_{ji}(I'_S), g_{ji}(I_T))}{\sum_{i}^{H_j, W_j, C_j} M'_i} \tag{2}$$

where d is either L_1 distance or L_2 distance squared, resulting in masked MAE or masked MSE perceptual loss functions.

Implicit Homography Estimation Loss (*iHomE*). Rather than directly minimizing only the distance between I_T and I'_S, we can encourage the network to simultaneously push I_T away from the original I_S in triplet fashion. The loss is highly inspired by Zhang *et al.* [44], but instead of learning the feature space to compare images, we use a fixed pre-trained *Loss Network* and non-learnable mask M_S. The iHomE loss can be formulated as:

$$L_{iHomE}(I_S, I_T) = \frac{\sum_{i}^{H_j, W_j} M'_i \cdot \max(ap_{ji} - an_{ji} + m, 0)}{\sum_{i}^{H_j, W_j} M'_i} \tag{3}$$

where m is triplet loss margin and ap_{ji} and an_{ji} are anchor-positive and anchor-negative channel-aggregated distances:

$$ap_{ji} = \frac{1}{C} \sum_{k}^{C_j} ||g_{jik}(I_T) - g_{jik}(I'_S)||_1$$

$$an_{ji} = \frac{1}{C} \sum_{k}^{C_j} ||g_{jik}(I_T) - g_{jik}(I_S)||_1 \tag{4}$$

Bidirectional implicit Homography Estimation Loss (*biHomE*). We can additionally leverage the fact that homography is invertible by estimating the transformation from the

target I_T to the source I_S images. Following Zhang *et al.* [44], we also add a constraint that enforces H_{ST} and H_{TS} to be inverse. We can formulate *biHomE* as:

$$L_{biHomE}(I_S, I_T) = L_{iHomE}(I_S, I_T) + \\ L_{iHomE}(I_T, I_S) + \\ \mu \|H_{ST}H_{TS} - \mathbb{1}\|_2^2 \tag{5}$$

where μ is the balancing hyperparameter [44] and $\mathbb{1}$ is 3×3 identity matrix.

4. Datasets

Since collecting the data for homography learning along with ground truth is hard, DeTone *et al.* [4] proposed to generate the dataset applying random projective transformations to COCO [20]. The Synthetic COCO (S-COCO) dataset was accepted by the research community, however, it does not model photometric distortion present in real-world images, such as contrast, brightness, and saturation changes. Here, we introduce a Photometrically Distorted Synthetic COCO (PDS-COCO) where we artificially model illumination changes by utilizing photometric distortion techniques used in [22] as augmentation for SSD object detector.

4.1. Synthetic COCO (S-COCO)

The dataset was introduced by DeTone *et al.* [4] in which source and target candidates are generated by duplicating the same COCO image [20]. The source patch I_S is generated by randomly cropping a source candidate at position p with a size of 128×128 pixels. Then the patch's corners are randomly perturbed vertically and horizontally by values within the range $[-\rho, \rho]$ and the four correspondences define a homography H_{ST}. The inverse of this homography $H_{TS} = (H_{ST})^{-1}$ is applied to the target candidate and from the resulted warped image a target patch I_T is cropped at the same location p. Both I_S and I_T are the input data with the homography H_{ST} as ground truth as shown in Figure 3. Such a procedure not only allows us to use large scale datasets but also creates ground truth homography labels for each transformation.

4.2. Photometrically Distorted Synthetic COCO (PDS-COCO)

To test the robustness of different homography estimation architectures to illumination changes we introduce Photometrically Distorted Synthetic COCO (PDS-COCO) dataset. The photometric distortion is inspired by Liu *et al.* [22] and the first step involves adjusting the brightness of the image using randomly picked value $\delta_b \in \mathcal{U}(-32, 32)$. Next, contrast, saturation and hue noise is applied with the following values: $\delta_c \in \mathcal{U}(0.5, 1.5)$, $\delta_s \in \mathcal{U}(0.5, 1.5)$ and $\delta_h \in \mathcal{U}(-18, 18)$. Finally, the color channels of the image
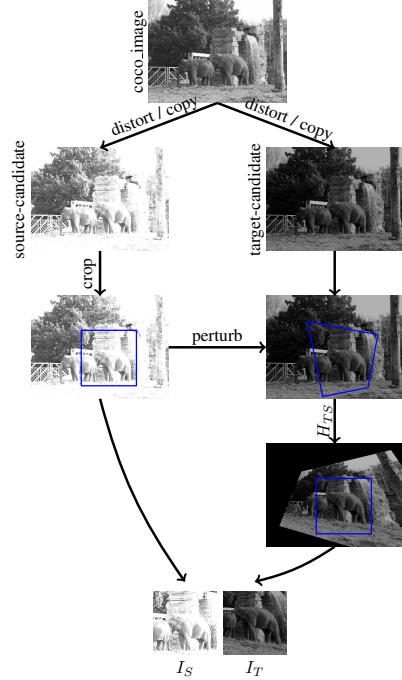

Figure 3. PDS-COCO dataset generation. Two separate photometric distortions are performed for a given COCO image resulting in the creation of source-candidate and target-candidate. Next, a random crop is selected for the source-candidate. The random crop is perturbed and given these correspondences, H_{ST} is computed. $H_{TS} = (H_{ST})^{-1}$ is applied to the target-candidate. Finally, source image I_S and target image I_T are extracted. Substituting photometric distortion with copy operation gives the original S-COCO dataset generation procedure.

are randomly swapped with a probability of 0.5. A composition of these distortions produces vastly different images, which constitutes a challenge for current homography estimation architectures. Such a photometric distortion procedure is applied to the original image independently to create source and target candidates. The rest of the procedure is the same as for S-COCO dataset generation as shown in the Figure 3.

5. Experiments

5.1. Experimental Setup

We implemented all the methods in PyTorch [30] and Kornia [32]. We modified the *HEN* architectures (if necessary) to use a ResNet34-like [14] feature extractor, to make sure the performance difference between methods comes from better design instead of a better CNN backbone. All methods are trained using the Adam optimizer [18] with a batch size of 64 and for $90k$ iterations. The initial learning rate (Detone and Nguyen: $5e^{-3}$, Zeng: $1e^{-3}$, Zhang: $1e^{-2}$) is decayed by a factor of 10 every $30k$ iterations. We use ResNet34 as a fixed *Loss Network* g with features taken

	Nguyen				Zhang				Zhang as *HEN + biHomE*			
$\delta=32$	NC	NC	NC	NC	0.89±0.24	NC	NC	NC	0.80±0.01	0.98±0.04	1.82±0.24	2.61±0.10
$\delta=24$	NC	NC	NC	NC	0.82±0.04	1.00±0.09	NC	NC	0.80±0.01	0.96±0.02	1.81±0.08	2.55±0.27
$\delta=16$	1.57±0.56	2.86±0.87	4.23±0.19	5.51±0.14	0.76±0.02	0.95±0.09	NC	NC	0.79±0.01	0.95±0.04	1.81±0.23	2.37±0.04
$\delta=8$	0.65±0.04	0.94±0.05	1.78±0.07	3.24±0.16	0.90±0.23	1.52±0.50	1.80±0.60	NC	0.84±0.06	0.95±0.04	1.60±0.06	2.20±0.13
$\delta=0$	0.48±0.02	0.79±0.07	1.25±0.05	2.07±0.10	0.74±0.01	0.95±0.01	1.67±0.61	2.61±0.81	0.79±0.01	0.97±0.06	1.16±0.11	1.87±0.07
	$\rho=8$	$\rho=16$	$\rho=24$	$\rho=32$	$\rho=8$	$\rho=16$	$\rho=24$	$\rho=32$	$\rho=8$	$\rho=16$	$\rho=24$	$\rho=32$

Figure 4. The viewpoint and illumination robustness comparisons of three unsupervised methods for different levels of viewpoint ρ and illumination δ changes. S-COCO is equivalent to the box with $\rho = 32$ and $\delta = 0$ and PDS-COCO is equivalent to the box with $\rho = 32$ and $\delta = 32$. NC stands for Not Converged (the methods which were not able to converge at least once per twenty independent runs). The method of Nguyen *et al*. [27] is robust to big viewpoint changes (big values of ρ) but is not able to converge for big illumination distortion (big values of δ). Zhang *et al*. [44] method is highly robust to illumination change δ but only for small viewpoint change ρ, whereas our *biHomE* loss is robust to both big illumination and viewpoint changes.

after the first residual block. We also used $N_S = 1024$ for Zeng method and $\mu = 0.01$, unless stated otherwise. The homography estimation quality is reported using Mean Absolute Corner Error (MACE) [4]. We report the mean and standard deviation of three runs.

5.2. S-COCO and PDS-COCO Results

Our implementation achieves much lower MACE on S-COCO than originally reported by DeTone *et al*. [4] and Nguyen *et al*. [27]: 1.96 and 2.07 against 9.20 and 12.91, respectively. This is because we employ a modern CNN architecture – ResNet34 [14] instead of originally used VGG-like structure [36]. Having the same backbone, Zeng's *et al*. [43] model is only about 13% better than DeTone's *et al*. [4] model – instead of 564% improvement reported in their paper.

As shown in the Figure 1, our architecture gives the best performance compared to all other unsupervised methods. Specifically, using our *biHomE* loss with the same *HEN* as Zhang *et al*. [44] demonstrates the beneficial effect of disentangling the homography estimation learning from feature representation learning. Moreover, our *biHomE* loss achieves comparable (Zeng *et al*. [43]) or better (DeTone *et al*. [4]) homography estimation performance than supervised methods on S-COCO.

Homography estimation on PDS-COCO presents a bigger challenge, because of additional photometric distortions on the images. Supervised approaches perform much worse in terms of MACE and neither Nguyen's [27] nor Zhang's [44] methods can converge. Our *biHomE* loss is the only method trained in an unsupervised manner that can converge on PDS-COCO. Similar to S-COCO, our *biHomE* loss achieves comparable or better performance compared to supervised methods.

5.3. Illumination and Viewpoint Robustness Study

The real-life images can exhibit big illumination and viewpoint changes. It is hard to collect such datasets with

ground truth homography labels, however, we can use the COCO dataset and simulate different conditions by two parameters: viewpoint change ρ and illumination change δ.

Viewpoint change ρ defines the maximum range of corner perturbation, and it was already introduced in Section 4.1. Since photometric distortion has a few elements, we bring them all together by illumination change δ so that all other deltas are controlled by one parameter:

$$\delta_x \in \mathcal{U}(-X \cdot \frac{\delta}{32}, X \cdot \frac{\delta}{32}) \qquad (6)$$

where δ_x is one of $\delta_b, \delta_c, \delta_s$ or δ_h (following the notation from Section 4.1), and X is the corresponding perturbation range. The bigger ρ the more viewpoint change and respectively the bigger δ the more illumination change between input images. For such a notation *S-COCO* is produced by $\rho = 32$ and $\delta = 0$ and *PDS-COCO* is produced by $\rho = 32$ and $\delta = 32$.

As presented in Figure 4, the method of Nguyen *et al*. [27] based on photometric loss is robust to big viewpoint changes (big values of ρ), but is not able to produce any reasonable model for big illumination distortion δ. On the other hand, Zhang *et al*. [44] method is highly robust to illumination change δ but only for small viewpoint change ρ. Our *biHomE* loss, which for a fair comparison we add on top of the same *HEN* as Zhang *et al*. [44] is robust both to big illumination δ and viewpoint ρ changes.

5.4. From Perceptual Loss to *biHomE*

Perceptual Loss was already introduced in the tasks of super-resolution [2, 19, 41], style transfer [16], image denoising [9], or training autoencoders [31]. The idea was to rely on perceptual similarity of high-level features extracted from pre-trained networks. The same property is desired in homography estimation. The architecture of *Loss Network* designed for super-resolution task, however, could be not optimal in the task of image comparison for homography estimation. To study the effects of different components of

	#1	#2	#3	#4	#5	#6	#7	#8
Loss network	VGG19	**ResNet34**	**ResNet34**	**ResNet34**	ResNet34	ResNet34	ResNet34	ResNet34
Layers	VGG22	Block2	**Block1**	Block0	Block1	Block1	Block1	Block1
Activation	Before	Before	Before	Before	**After**	After	After	After
Loss function	MSE	MSE	MSE	MSE	MSE	L_1	**iHomE**	**biHomE**
MACE	**8.59** ±0.05	7.15 ±0.16	**4.18** ±0.16	4.74 ±0.05	**4.11** ±0.11	3.82 ±0.11	3.67 ±0.10	**2.39** ±0.20

Table 1. Comparison of performance on PDS-COCO of different settings of perceptual loss. Each column represents a model with DeTone's [4] architecture as HEN and a particular perceptual loss configuration. The red color indicates the main improvement compared with the previous model starting from Wang *et al.* [41] setting.

the perceptual loss we gradually modified the State-of-the-Art perceptual loss proposed by Wang *et al.* [41] and compare their performance in the homography estimation task as presented in Table 1. The red color indicates the main improvement compared with the previous model. We use DeTone *et al.* [4] in all configurations and only change elements of perceptual loss. The average MACE of three runs is presented. A detailed discussion is provided as follows.

ResNet34. We first move from VGG19 [36] network to ResNet34 [14]. Using ResNet as a *Loss Network* significantly improves the homography estimation accuracy, which can be caused by a better-learned feature representation. We also noticed that learning is more likely to converge with ResNet, probably due to better gradient flow of residual connections. The detailed comparison is out of the scope of this article. For a comprehensive analysis of these networks please refer to [1, 17].

Block1. Ledig *et al.* [19] argues to use higher-level features from deeper network layers with the potential to focus more on the content of the image. It is essential for the Super-Resolution task, but as depicted in Table 1, in Homography Estimation low-level features matter more.

Features after activation. In contrast to Wang *et al.* [41], we show that it is beneficial to use features after the ReLU activation function [10]. A possible explanation is that low-level features are activated mostly on edges and corners, which are crucial for image alignment purposes (please refer to Figure 6a of Wang *et al.* [41] for a visual comparison).

HEN	Zhang	Zhang	Zhang
Loss Function	Triplet Loss	*biHomE*	*biHomE*
Channel-	-	c-wise	c-agnostic
$m = 1$	3.07 ±0.34	4.57 ±0.21	3.33 ±0.34
$m = 100$	2.50 ±0.25	2.84 ±0.27	1.86 ±0.16
$m = \infty$	**2.08** ±0.11	**1.87** ±0.07	**1.87** ±0.07

Table 2. The performance on the S-COCO dataset of Zhang *et al.* [44] and our *biHomE* loss with different margin m settings and different channel information aggregation. The *infinite* margin version is consistently better for both losses and it is used in all experiments in the paper.

L_1 **distance.** Replacing MSE loss function with L_1 distance yields better homography estimates. Due to warping and illumination differences, feature maps could contain some artifacts, and L_1 loss is known to be more robust to outliers [11].

iHomE. Triplet loss introduced by Zhang *et al.* [44] can further improve the homography estimation performance. This is probably because the model has additional information of original $I1$ image features.

biHomE. Using triplet loss in both directions similarly to Zhang *et al.* [44] can additionally improve the results. The improvement is even more profound for other *HENs*.

5.5. Margin Ablation

Zhang *et al.* [44] used a margin value of $m = 1$ in their triplet loss formulation. However, we found out that on S-COCO bigger values of m yield better results (Table 2). The best performance is produced when the triplet loss is never saturated, which is equivalent to setting the margin to an *infinite* value:

$$L_j^{iHomE}(I_S, I_T) = \frac{\sum_{i}^{H_j W_j} M_i' \cdot (ap_{ji} - an_{ji})}{\sum_{i}^{H_j W_j} M_i'} \qquad (7)$$

In all the experiments for both Zhang *et al.* [44] and our iHomE and *biHomE* losses we use above Formulation 7 instead of Equation 3.

5.6. Channel Information Aggregation

In contrast to Zhang *et al.* [44], our feature map obtained by *Loss Network g* has more than one channel, thus we have to figure out how the channel information will be aggregated. One can either calculate distances for every channel and then apply triplet loss in a channel-agnostic way per every spatial location or apply triplet per every channel location (channel-wise). The former is presented in equations

3 and 4, and the latter can be formulated as:

$$L_j^{iHomE}(I_S, I_T) = \frac{\sum_{i}^{H_j W_j C_j} M_i' \cdot \max(ap_{ji} - an_{ji} + m, 0)}{\sum_{i}^{H_j W_j C_j} M_i'}$$

(8)

where;

$$ap_{ji} = ||g_{ji}(I_T) - g_{ji}(I_S')||_1$$
$$an_{ji} = ||g_{ji}(I_T) - g_{ji}(I_S)||_1$$

(9)

As presented in Table 2 the channel-agnostic way achieves better performance on the S-COCO dataset. We hypothesize that forcing every channel of I_T feature description to be closer to I_S' than to I_S is probably a very hard task and looking at all channels before comparison makes the optimization easier. Channel-agnostic and channel-wise versions of *biHomE* loss with *infinite* margin are mathematically equivalent, so we report only one number.

5.7. To Freeze or not to Freeze

We also want to find out how important is freezing the *Loss Network* for effective homography estimation learning. As shown in Table 3, fixed *Loss Network* performs better. One of the possible reasons is that freezing weights of g allows using a bigger learning rate, which can result in better convergence [37, 38]. But even for the same learning rate value freezing the *Loss Network* is still a preferable policy.

	lr	S-COCO MACE
Zhang+*biHomE*	1e-2	**1.87** ±0.07
Zhang+*biHomE*	1e-3	2.03 ±0.04
Zhang+*biHomE* (learnable)	1e-3	2.17 ±0.13

Table 3. S-COCO performance of Zhang *et al.* [44] as *HEN* and our *biHomE* loss when using fixed and learnable *Loss Network*. Fixing weights of g allows using bigger learning rate, which can result in better convergence [37, 38].

5.8. *biHomE* Performance on Out-of-Distribution Dataset

One of the shortcomings of using models pre-trained on ImageNet [35] as *Loss Network* is possible sub-optimal results on out-of-distribution datasets. To test this vulnerability we compare Zhang *et al.* [44] method with the original and our *biHomE* loss on FLIR Dataset [7] preprocessed similar to S-COCO. As shown in Table 4, although thermal images are vastly different than color images the *Loss Network* was trained on, our *biHomE* loss still achieves better performance.

	S-FLIR MACE
Zhang+*biHomE*	**1.59** ±0.13
Zhang+TripletLoss	2.08 ±0.25

Table 4. The performance of our *biHomE* loss vs. original Zhang *et al.* [44] method on S-FLIR dataset [7] containing thermal images for ADAS systems.

6. Image Alignment for Change Captioning

In this Section, we study the influence of image alignment using homography on change captioning task. The goal is to generate captions for an image pair before and after the change. There are six change types (color/material change, adding/dropping/moving an object, and no change) and each is combined with illumination/viewpoint change. Recently, Park *et al.* [28] proposed a new architecture called Dual Dynamic Attention Model (DUDA) for generating change captions. It entails three main components: feature extractor (ResNet101 [14]), Dual Attention and Dynamic Speaker modules. The Dual Attention module takes features produced by the feature extractor and generates a sparse feature vector with a description of the changes between input images. The Dynamic Speaker module then produces the caption word-by-word dynamically moving the attention between the first and second images depending on the stage of the sentence.

In their evaluation Section, Park *et al.* [28] showed degradation of captioning performance as viewpoint shift increases – our reproduced results are shown in Figure 5. The Dual Attention module intrinsically assumes the im-
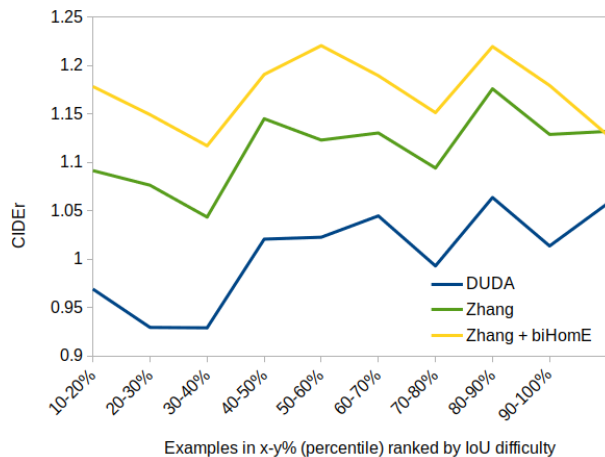

Figure 5. Change captioning performance breakdown by viewpoint shift (measured by IoU). Aligning images with Zhang [44] method before learning to generate captions improves the metrics, but performance degradation between smaller and bigger viewpoint shifts is still present. Learning Zhang [44] with our *biHomE* loss seems to alleviate this problem.

Approach	Total				Scene Change				Distractor			
	B	C	M	S	B	C	M	S	B	C	M	S
{No alignment} + DUDA [28]	53.0	115.2	37.4	31.3	50.8	100.3	33.2	27.8	62.3	115.9	49.8	34.8
{Zhang [44] + TripletLoss} + DUDA [28]	**54.5**	120.7	**40.2**	**33.0**	**52.1**	111.4	36.0	**31.4**	61.2	115.3	**51.6**	34.6
{Zhang [44] + *biHomE*} + DUDA [28]	53.0	**124.5**	**40.2**	**33.0**	50.6	**117.3**	**36.9**	30.9	**63.7**	**117.1**	50.5	**35.0**

Table 5. Change Captioning evaluation on CLEVR_Change dataset. Aligning images before learning to generate captions improves all captioning metrics: BLEU-4 (B), CIDEr (C), METEOR (M), and SPICE (S) in each setting (i.e. Total, Scene Change, Distractor). For more details, we refer to Park *et al*. [28]. The model trained with our *biHomE* loss achieves the best result in most of the cases.

ages before and after the change are roughly aligned, thus finding corresponding objects on images with big viewpoint change is challenging. We argue that aligning those images helps DUDA architecture learn better captions of the change on the scene.

We conducted experiments for Zhang *et al*. [44] architecture with and without our *biHomE* loss. Unless stated otherwise, the following are the steps that have been followed in all our experiments: first, the homography estimation is learned. Second, we transform images after change onto images before change using the trained homography estimation model. The pre-processed images are trained using the DUDA model and the captioning quality metric is reported in Table 5. Moreover, we also prepare similar validation of the robustness of the model to viewpoint shift, according to the methodology shown in Park *et al*. [28] which is depicted in Figure 5.

Experimental evaluation shows that aligning images before learning to caption is beneficial not only for overall captioning quality but also for performance degradation between smaller and bigger viewpoint shifts. A sample image

before change   after change

DUDA

the small brown metal cylinder that is in front of
the small purple rubber object has been newly placed
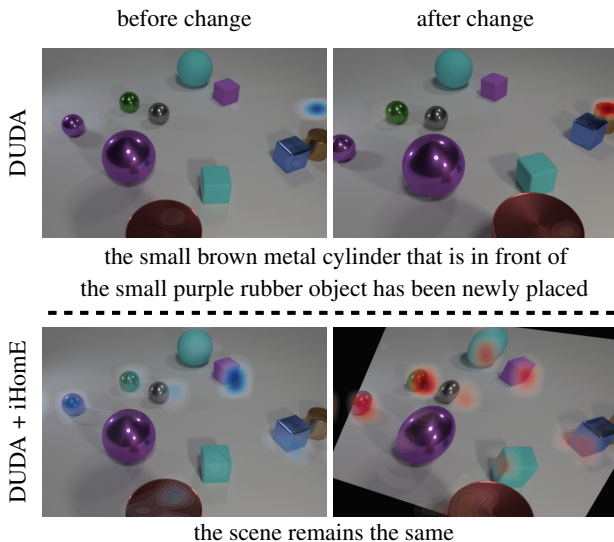
DUDA + iHomE

the scene remains the same

Figure 6. Qualitative result comparing DUDA [28] output with and without image alignment. The blue and red attention maps are applied before and after the change, respectively. After adding the image alignment step, DUDA [28] architecture produces the correct caption.

pair with caption generated with and without image alignment step is presented in Figure 6. Note that using our *bi-HomE* loss the performance degradation is negligible for the CLEVR_Change dataset, which can be explained by facilitated correspondence search in the Dual Attention module on aligned images.

7. Conclusion

We presented an unsupervised approach to homography estimation that is robust to big illumination and viewpoint changes at the same time. We showed that disentangling the homography estimation from representation learning provides better estimates. We also proposed to use an additional photometric distortion step in the synthetic COCO dataset generation (PDS-COCO) and encourage future works to use it as a new evaluation benchmark of robust homography estimation. Then, we presented a study of modern homography estimation baselines along with our bidirectional implicit Homography Estimation (*bi-HomE*) loss on both S-COCO and PDS-COCO. *biHomE* loss achieves a new state-of-the-art performance for unsupervised homography estimation, which is also comparable or better compared to supervised approaches. Furthermore, we investigated the influence of image alignment using homography on change captioning quality. We showed that aligning the images using our *biHomE* loss is not only beneficial for the overall captioning quality of DUDA architecture on the CLEVR_Change dataset but also for performance degradation from smaller to bigger viewpoint shifts.

References

[1] Simone Bianco, Remi Cadene, Luigi Celona, and Paolo Napoletano. Benchmark Analysis of Representative Deep Neural Network Architectures. *IEEE Access*, 6:64270–64277, 2018. 6

[2] Joan Bruna, Pablo Sprechmann, and Yann LeCun. Super-Resolution with Deep Convolutional Sufficient Statistics. In *4th International Conference on Learning Representations, ICLR 2016*, 2016. 5

[3] Ryan L Buffington and John E McInroy. Homography-based change detection for space-based satellite inspection. In *Sensors and Systems for Space Applications IV*, volume 8044, page 804409. International Society for Optics and Photonics, 2011. 1

[4] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep Image Homography Estimation. *arXiv preprint arXiv:1606.03798*, 2016. 1, 2, 3, 4, 5, 6

[5] Jakob Engel, Thomas Schöps, and Daniel Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *European Conference on Computer Vision*, pages 834–849. Springer, 2014. 1

[6] Georgios D Evangelidis and Emmanouil Z Psarakis. Parametric Image Alignment Using EnhancedCorrelation Coefficient Maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1858–1865, 2008. 2

[7] F. A. Group. FLIR Thermal Dataset for Algorithm Training. https://www.flir.in/oem/adas/adas-dataset-form. accessed 26.02.2021. 7

[8] 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

[9] Maryam Gholizadeh-Ansari, Javad Alirezaie, and Paul Babyn. Deep Learning for Low-Dose CT Denoising Using Perceptual Loss and Edge Detection Layer. *Journal of Digital Imaging*, pages 1–12, 2019. 5

[10] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep Sparse Rectifier Neural Networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011. 6

[11] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016. 6

[12] Heng Guo, Shuaicheng Liu, Tong He, Shuyuan Zhu, Bing Zeng, and Moncef Gabbouj. Joint Video Stitching and Stabilization From Moving Cameras. *IEEE Transactions on Image Processing*, 25(11):5491–5503, 2016. 1

[13] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003. 1, 2, 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3, 4, 5, 6, 7

[15] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2462–2470, 2017. 1

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 1, 2, 5

[17] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the Recent Architectures of Deep Convolutional Neural Networks. *Artificial Intelligence Review*, 53(8):5455–5516, 2020. 6

[18] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[19] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017. 5, 6

[20] 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. 4

[21] Shuaicheng Liu, Ping Tan, Lu Yuan, Jian Sun, and Bing Zeng. Meshflow: Minimum latency online video stabilization. In *European Conference on Computer Vision*, pages 800–815. Springer, 2016. 1

[22] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single Shot MultiBox Detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016. 4

[23] David G Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 2

[24] Bruce D Lucas and Takeo Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence – Volume 2*, pages 674–679, 1981. 2

[25] Simon Lucey, Rajitha Navarathna, Ahmed Bilal Ashraf, and Sridha Sridharan. Fourier Lucas-Kanade Algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1383–1396, 2012. 2

[26] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 1

[27] Ty Nguyen, Steven W Chen, Shreyas S Shivakumar, Camillo Jose Taylor, and Vijay Kumar. Unsupervised Deep Homography: A Fast and Robust Homography Estimation Model. *IEEE Robotics and Automation Letters*, 3(3):2346–2353, 2018. 1, 2, 3, 5

[28] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust Change Captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4624–4633, 2019. 2, 7, 8

[29] Seonwook Park, Thomas Schöps, and Marc Pollefeys. Illumination Change Robustness in Direct Visual SLAM. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 4523–4530. IEEE, 2017. 1

[30] 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, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 4

[31] Gustav Grund Pihlgren, Fredrik Sandin, and Marcus Liwicki. Improving Image Autoencoder Embeddings with Perceptual Loss. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020. 5

[32] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an Open Source Differentiable Computer Vision Library for PyTorch. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3674–3683, 2020. 4

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 2

[34] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *2011 International Conference on Computer Vision*, pages 2564–2571. IEEE, 2011. 2

[35] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 7

[36] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 5, 6

[37] Leslie N Smith. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018. 7

[38] Leslie N Smith and Nicholay Topin. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, page 1100612. International Society for Optics and Photonics, 2019. 7

[39] Richard Szeliski. Image alignment and stitching: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 2(1):1–104, 2006. 1, 2

[40] Dennis WJM van de Wouw, Gijs Dubbelman, and Peter HN de With. Hierarchical 2.5D Scene Alignment for Change Detection with Large Viewpoint Differences. *IEEE Robotics and Automation Letters*, 1(1):361–368, 2016. 1

[41] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2, 5, 6

[42] Fuli Wu and Xianyong Fang. An improved RANSAC homography algorithm for feature based image mosaic. In *Proceedings of the 7th WSEAS International Conference on Signal Processing, Computational Geometry & Artificial Vision*, volume 2. World Scientific and Engineering Academy and Society (WSEAS), 2007. 2

[43] Rui Zeng, Simon Denman, Sridha Sridharan, and Clinton Fookes. Rethinking Planar Homography Estimation Using Perspective Fields. In *Asian Conference on Computer Vision*, pages 571–586. Springer, 2018. 1, 2, 3, 5

[44] Jirong Zhang, Chuan Wang, Shuaicheng Liu, Lanpeng Jia, Nianjin Ye, Jue Wang, Ji Zhou, and Jian Sun. Content-Aware Unsupervised Deep Homography Estimation. *arXiv preprint arXiv:1909.05983*, 2019. 1, 2, 3, 4, 5, 6, 7, 8