

Hala Lamdouar, Weidi Xie, Andrew Zisserman.

Visual Geometry Group VGG
Department Of Engineering Science
University of Oxford

Motivation:

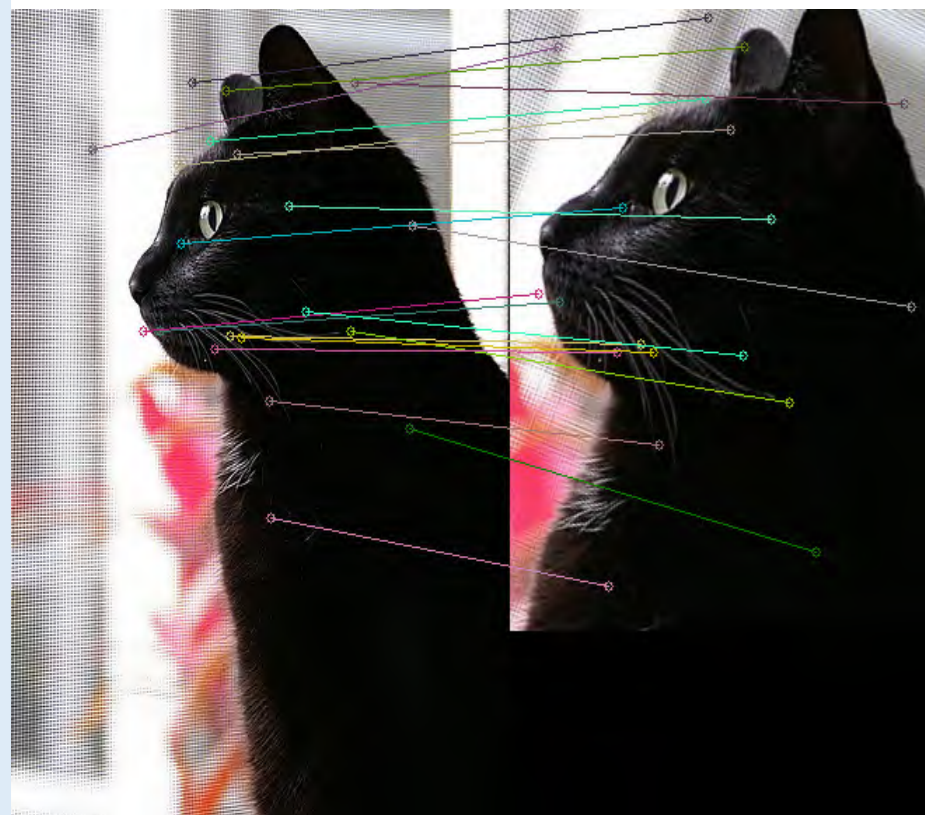


Image alignment is a classical task that involves finding correspondences and inferring geometric transformation that maps a given source image to a target image.

- Current methods lack robustness and fail to align the complex and ambiguous cases.

Contributions:

- Unsupervised **robust** alignment network that emphasises relevant features and handles outliers.
- **Burstiness** module that penalises the repetitive regions or the textures within the image.

Self-Burstiness:

Consider an image $I \in \mathcal{R}^{M \times N}$ and its feature representation $\mathcal{F}_I \in \mathcal{R}^{M \times N \times D}$

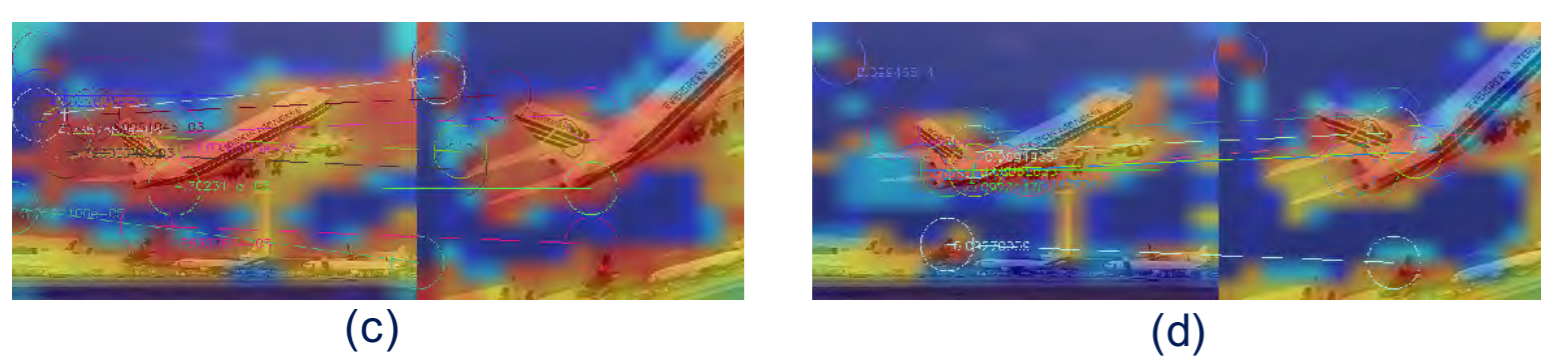
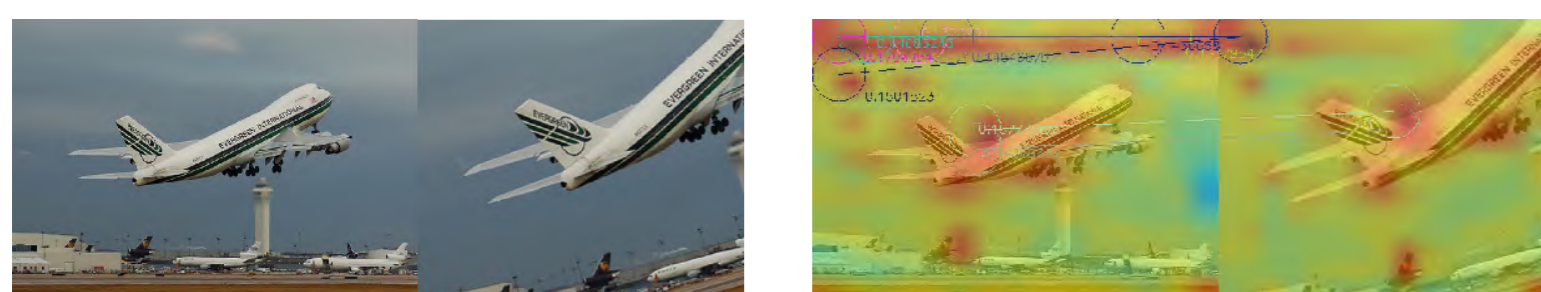
For $(i, j) \in I$
 $U_{i,j}^I = \text{softmax}(-\mathcal{B}_r^I(f_{i,j}))$

Repetitiveness of the feature $f_{i,j}$ within the open ball $\mathcal{B}_r^I(f_{i,j})$

$$\mathcal{B}_r^I(f_{i,j}) = \{f_{k,l} \in \mathcal{F}_I \mid \|f_{i,j} - f_{k,l}\|_2 < r\}$$

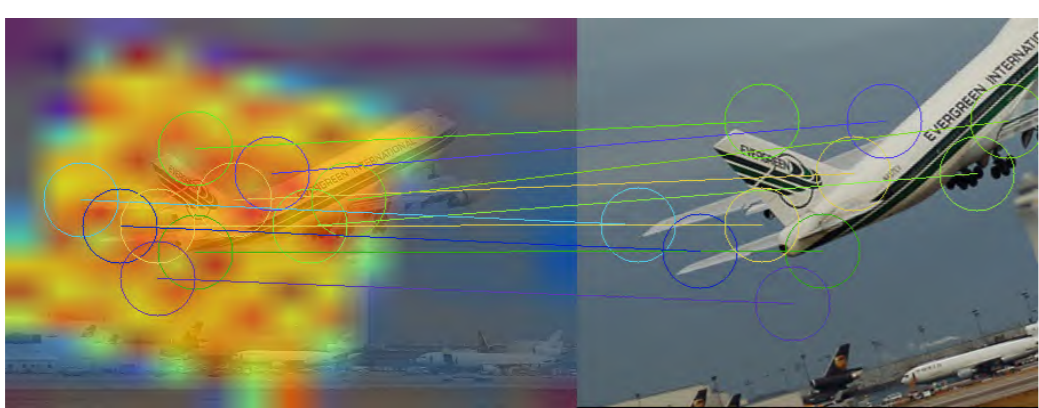
- Burstiness-based matching score for each $(i, j) \in I_{src}, (k, l) \in I_{tgt}$

$$S_{i,j,k,l}^{I_{src}, I_{tgt}} = \langle f_{i,j} \odot U_{i,j}^{I_{src}}, f_{k,l} \odot U_{k,l}^{I_{tgt}} \rangle$$



(a): original pair of images with the source (left) and the target (right), (b): the best matching scores and the top correspondences without self-burstiness, in (c) with self-burstiness $r = 0.9$ and (d) with self-burstiness $r = 0.8$

RANSAC-like inlier mask:



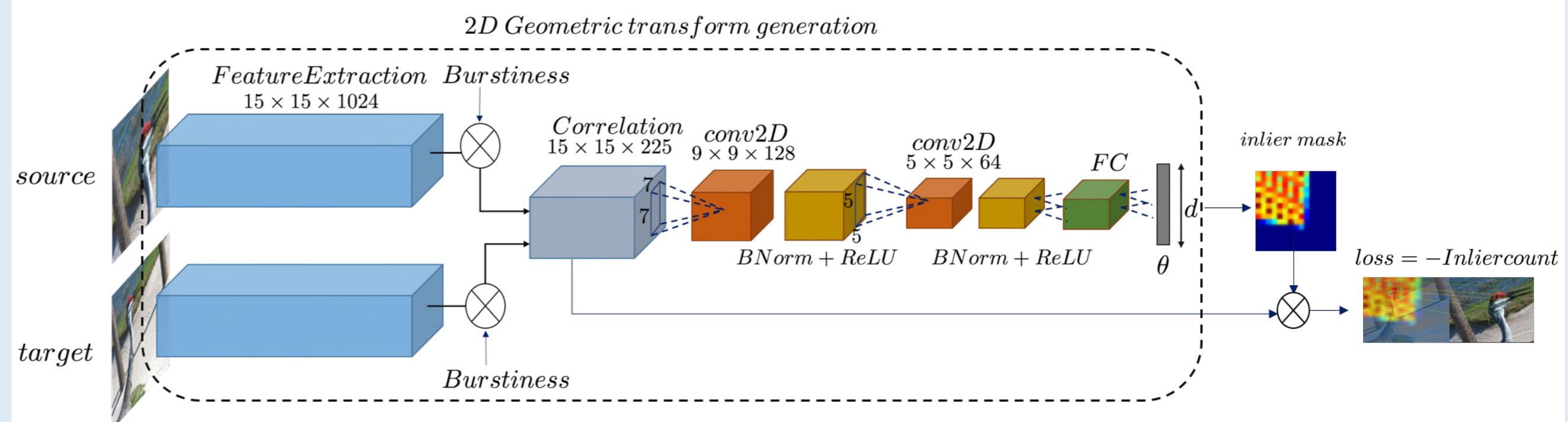
For each $(i, j) \in I_{src}, (k, l) \in I_{tgt}$

$$m_{i,j,k,l} = \begin{cases} 1 & \text{if } \|(i, j) - \mathcal{T}^{-1}_\theta(k, l)\|_2 < l \\ 0 & \text{otherwise} \end{cases}$$

Model:

End-to-end alignment network including:

- A pre-trained feature extraction convolutional neural network (VGG-16, ResNet-101)
- A burstiness module (downweights irrelevant features)
- 2D geometric transform generation (convolutional regression network)
- A differentiable RANSAC-like inlier count (inlier mask)

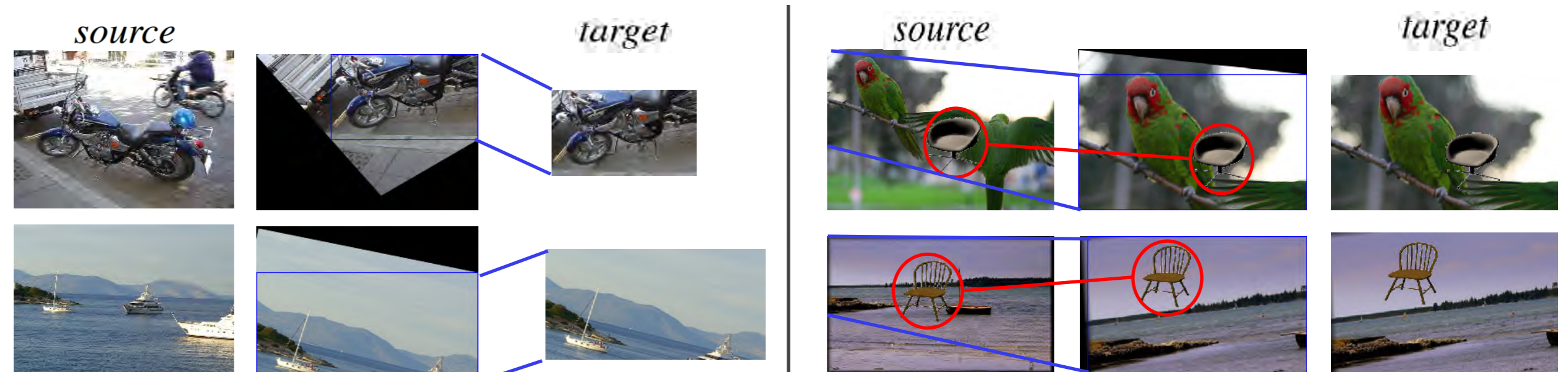


Loss function:

$$\mathcal{L}_{inliers} = - \sum_{(i,j) \in I_{src}} \sum_{(k,l) \in I_{tgt}} S_{i,j,k,l}^{I_{src}, I_{tgt}} m_{i,j,k,l}$$

Challenging dataset:

- 2 different transformations for the background (PASCAL VOC2011) and the foreground (3D rendered chairs)
- Cropping with respect to the maximal inscribed axis-aligned rectangle



Results:

We use the PCK measure (percentage of correct points) i.e. the number of points where:

$$\|\mathcal{T}_\theta(i, j) - \mathcal{T}_{\theta_{GT}}(i, j)\|_2 < l$$

On PF-PASCAL dataset:

class/method	VGG-16-affine	ResNet-101-affine	VGG-16-affine+TPS	ResNet-101-affine+TPS	VGG-16-affine+Burstiness	ResNet-101-affine+Burstiness	VGG-16-affine+TPS+Burstiness	ResNet-101-affine+TPS+Burstiness
acroplane	50.40%	50.98%	50.37%	44.53%	50.76%	44.92%	51.15%	18.58%
bicycle	42.61%	43.23%	43.20%	38.50%	43.49%	39.30%	41.41%	12.90%
bird	31.67%	28.26%	36.44%	30.76%	32.47%	32.70%	34.62%	20.88%
boat	27.78%	30.56%	26.85%	19.44%	22.22%	33.33%	25.00%	23.15%
bottle	39.17%	10.00%	37.92%	16.25%	47.08%	42.08%	46.25%	8.46%
bus	49.07%	41.65%	49.85%	37.48%	51.28%	46.26%	52.33%	8.46%
car	51.51%	43.32%	52.83%	42.71%	48.36%	38.73%	48.14%	12.53%
cat	30.30%	28.05%	29.33%	34.04%	30.20%	29.51%	30.04%	25.47%
chair	36.35%	25.24%	37.54%	27.38%	33.49%	38.65%	35.63%	13.41%
cow	65.56%	54.44%	65.56%	65.56%	71.11%	52.22%	71.11%	12.22%
diningtable	33.93%	28.57%	32.14%	25.60%	39.29%	28.57%	35.71%	16.07%
dog	29.05%	27.94%	28.35%	30.17%	27.94%	32.46%	29.57%	22.67%
horse	30.18%	22.00%	30.18%	21.47%	33.72%	27.24%	31.87%	21.61%
motorbike	41.62%	40.90%	40.86%	35.77%	42.33%	39.40%	40.51%	12.00%
person	26.65%	24.23%	28.50%	26.13%	28.76%	25.21%	26.95%	17.36%
pottedplant	31.04%	28.96%	33.54%	28.96%	31.67%	22.29%	34.79%	18.33%
sheep	80.00%	40.00%	80.00%	80.00%	80.00%	60.00%	80.00%	0.00%
sofa	37.82%	47.65%	39.25%	32.61%	39.58%	37.04%	38.78%	14.17%
train	40.00%	40.00%	39.00%	36.00%	39.00%	31.00%	39.00%	5.00%
tvmonitor	20.11%	13.00%	22.11%	14.89%	27.11%	27.67%	29.56%	1.11%
total	38.26%	34.55%	38.64%	33.06%	39.10%	35.96%	39.17%	14.35%

References:

- [1] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In Advances in neural information processing systems, pages 2017–2025, 2015.
- [2] I. Rocco, R. Arandjelovic, and J. Sivic. End-to-end weakly supervised semantic alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6917–6925, 2018.
- [3] I. Rocco, R. Arandjelovic, and J. Sivic. Convolutional neural network architecture for geometric matching. In Proc. CVPR, volume 2, 2017.
- [4] D. G. Lowe. Distinctive image features from scale invariant keypoints. International journal of computer vision, 60(2):91–110, 2004.
- [5] H. Jegou, M. Douze, and C. Schmid. On the burstiness of visual elements. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 1169–1176. IEEE, 2009.
- [6] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother. Dsac-differentiable ransac for camera localization. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), volume 3, 2017.