Supplementary Material for the Paper: Unsupervised Homography Estimation with Coplanarity-Aware GAN

Mingbo Hong[1,2*] Yuhang Lu[1,3*] Nianjin Ye[1] Chunyu Lin[4] Qijun Zhao[2†] Shuaicheng Liu[5,1†]
[1] Megvii Technology [2] Sichuan University [3] Univesity of South Carolina
[4] Beijing Jiaotong University [5] University of Electronic Science and Technology of China

1. Ablation study on the GAN loss

In Eq. (5) of the paper, we employ a hybrid loss function with three terms to optimize the coplanarity-aware GAN. To provide an in-depth understanding of the proposed GAN, we design a series of ablation experiments to analyze the usefulness of each loss term. The results are reported in Table 1.

In row 2, we do not use any part of the proposed GAN or the plane masks in our method, and obtain an average error of 0.49. In row 3, we add the generator back to the network, and use the auxiliary loss L_{aux} to contrast the generated masks with a constant mask. The generated masks also participate in the calculation of the alignment loss L_{align}, such that they can be implicitly optimized together with the predicted homographies. The auxiliary loss L_{aux} helps prevent the generated masks from being all-zero. The comparison between row 2 and row 3 validates the effectiveness of mask prediction in homography estimation, with the average error decreasing from 0.49 to 0.44. Even though the masks are optimized without explicit guidance, they can still learn to exclude some disturbance or uninformative regions.

In row 4, we add the whole coplanarity-aware GAN back and begin to use the adversarial loss L_{adv} in training. Comparing with row 3, we can see that the average error decreases by only 0.01. But it does not mean the adversarial loss is not helpful. The adversarial loss needs to be regularized by the gradient penalty L_{gp} to properly train the generator and discriminator. When the gradient penalty is also used (row 7), the average error can be further reduced to 0.39. These experiments verify the effectiveness of L_{adv} and L_{gp} as a whole. With the coplanarity constraint provided by L_{adv}, the masks are able to focus on the dominant plane only. In row 5, we try to remove the auxiliary loss L_{aus} from our method. However, we find that the generator can easily collapse to all-zero solutions in this case, and the network training does not converge. It indicates that using a constant mask for regularization is necessary.

As mentioned in Section 3.3 of the paper, we employ the gradient reversal layer [1] to facilitate one-stage adversarial training. In row 6, we discard the gradient reversal layer, and train the proposed GAN with a regular two-stage training strategy, which alternately fixes the generator or the discriminator and updates the other one. Comparing with row 7, we can see that the one-stage strategy does not only help to improve the performance of our method, but also make the training more convenient.

2. Visualization of the weight token

In Section 3.2 of the paper, a weight token is employed in the decoding stage to summarize useful information from self-attention features to predict the weights of 8 flow bases. To better understand the role of the weight token, we visualize it by averaging on the channel dimension. For each input patch of size 384×512, we use our trained model to obtain an attention map of size 48×64, and then resize it to the input size. The visualization results are illustrated in Fig. 1. We can see that, after jointly optimizing with the coplanarity-aware GAN, the weight token in the transformer is able to focus on regions in the dominant plane by itself, thus making the transformer predict plane induced homographies in testing.

3. Visualization of plane masks

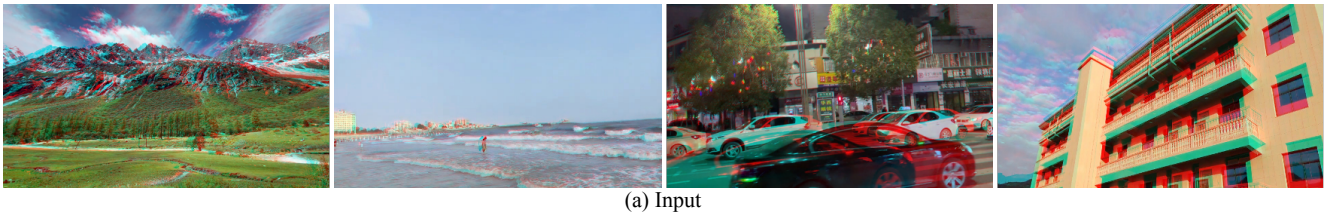
In Fig. 4, we provide additional visualization results of our masks in several challenging scenarios, including low light, small foreground, large foreground, etc. Besides, we also provide the masks of CA-Unsupervised [2] for comparison. Comparing to our masks, CA masks tend to focus on texture-rich regions rather than the dominant plane. Without explicit constraint, CA-Unsupervised implicitly learns to focus on regions with rich textures, and suppress unregisterable regions such as moving objects. However, with the coplanarity constraint, our method explicitly optimizes the mask generator to focus on the dominant plane, thus making the mask even cleaner and further improving the homography estimation. For example, in Fig. 4(a), although CA masks can pay attention to the buildings, they do not re-

*Equal contribution. †Corresponding authors.

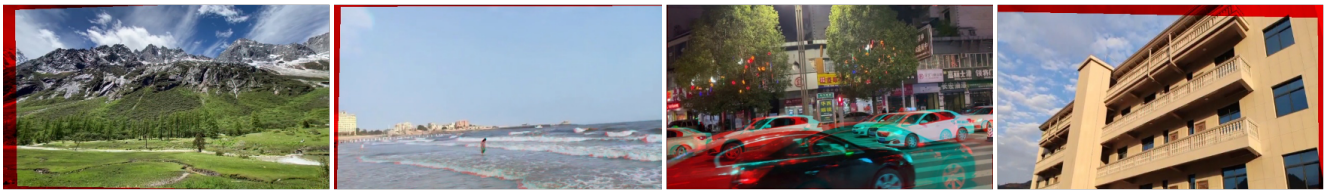
1)	L_{adv}	L_{gp}	L_{aux}	Adversarial training	RE	LT	LL	SF	LF	Avg
2)	-	-	-	-	0.26(+18.18%)	0.59(+43.90%)	0.59(+3.51%)	0.63(+43.18%)	0.40(+29.03%)	0.49(+25.64%)
3)	-	-	✓	-	0.24(+9.09%)	0.50(+21.95%)	0.64(+12.28%)	0.59(+34.09%)	0.36(+16.13%)	0.44(+12.82%)
4)	✓	-	✓	one-stage	0.23(+4.55%)	0.55(+34.15%)	0.59(+3.51%)	0.46(+4.55%)	0.32(+3.23%)	0.43(+10.26%)
5)	✓	✓	-	one-stage	-	-	-	-	-	-
6)	✓	✓	✓	two-stage	0.23(+4.55%)	0.44(+7.32%)	0.60(+5.26%)	0.44(+0.00%)	0.31(+0.00%)	0.40(+2.56%)
7)	✓	✓	✓	one-stage	0.22(+0.00%)	0.41(+0.00%)	0.57(+0.00%)	0.44(+0.00%)	0.31(+0.00%)	0.39(+0.00%)

Table 1. Results of ablation experiments on the proposed GAN loss L_{plane} in Eq. (5) of the paper.

Figure 1. Visualization of the weight token.



(a) Input



(b) Output

Figure 2. Alignment results of our method on unseen scenarios.

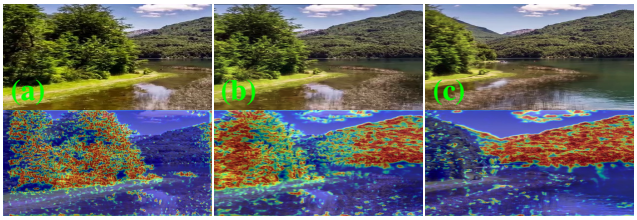


Figure 3. Scenes with multiple candidate planes.

move the cars thoroughly, especially at the car boundaries, which will downgrade the accuracy of the predicted homography. In contrast, our masks could highlight the building region only.

In addition to the coplanarity constraint, there is another difference that makes us produce better masks than CA-Unsupervised. In CA-Unsupervised, the masks are predicted from only one of the source and target images. However, our masks are always predicted from a pair of source and target images, which enables our mask generator to incorporate the information from both images to identify unregisterable regions.

Our method is also able to handle scenes multiple candidate planes. For example, in Fig. 3, we show three examples of the same scene with 4 planes, *i.e.*, sky, lake, forest and mountain. From Fig. 3 (a) to (c), the camera is gradu-

ally moving from left to right, and the corresponding mask switches from the forest to the mountain. When the forest and mountain have similar areas in (b), our method will compromise to highlight both planes.

4. Generalization

To further examine the generalization ability of the proposed method, we test our model in several different scenarios that are unseen in the training set. In Fig. 2, we display our results on several unseen scenarios, including mountain, sea, street view, and buildings. Despite that these images have different feature distributions from the training images, our model still works well on aligning their dominant planes. It indicates the potential of our method to be deployed to practical applications.

References

[1] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 1

[2] Jirong Zhang, Chuan Wang, Shuaicheng Liu, Lanpeng Jia, Nianjin Ye, Jue Wang, Ji Zhou, and Jian Sun. Content-aware unsupervised deep homography estimation. In *Proc. ECCV*, pages 653–669, 2020. 1

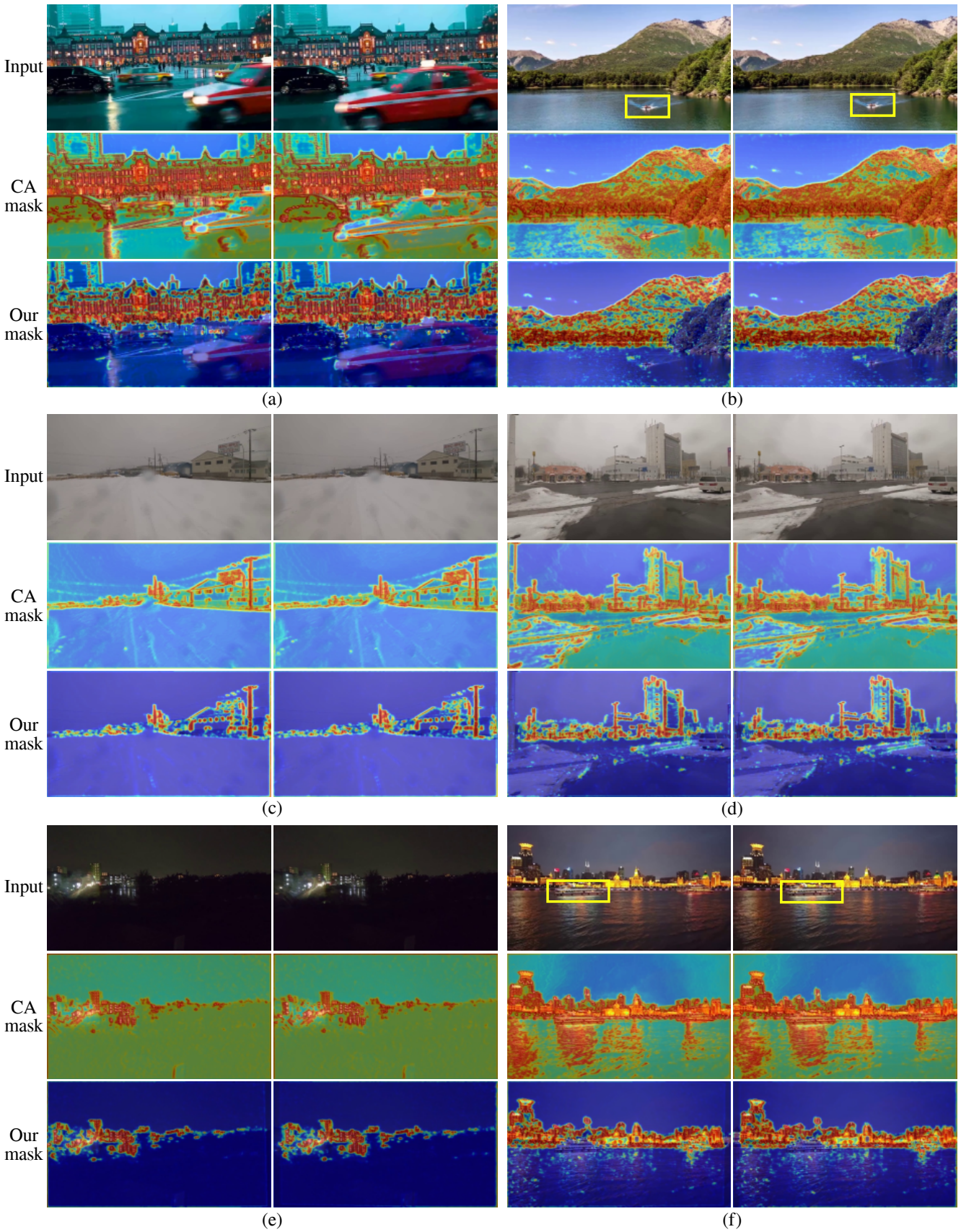


Figure 4. Visualization of CA masks and our masks.