

# SSSS: Synthesized Sequence Stereo Scenes

謝欣玉 施君諺 黃秉茂  
Group-05 R11922069 R11922158 R11944024

## Abstract

我們參考了數篇論文的內容並結合我們自己的方法，提出了一個能將單目序列影像合成出雙目序列影像的完整流程。以此流程生成的雙目序列影像，可以用來執行 SLAM，產生和真實地圖相近的里程計地圖。

## 1 Introduction

在上課時，我們學到，可以利用雙目相機拍攝序列影像去執行 SLAM 生成里程計地圖，但這樣需要特地買一臺雙目相機。在沒有額外經費的情況下，一般來說，我們手邊只有單目鏡頭的裝置，例如：手機、單目相機等。此外，雙目鏡頭裝置建置成本較單目鏡頭裝置高，若能從單目影像生成雙目影像，不只能降低建置成本，還可以改善單目鏡頭裝置的尺度不確定性等問題。因此我們嘗試只使用單目相機拍攝序列影像，利用單目序列影像合成出雙目序列影像，並且在執行 SLAM 後，也能產生和真實地圖相近的里程計地圖。我們的模型使用 SceneFlow Dataset 訓練，並使用 KITTI Dataset 執行 SLAM，產生里程計地圖和進行 evaluation。

## 2 Related Work

在 Watson et al.[1] 的 work 中，其提出以下的 pipeline，將合成雙目序列影像的步驟分成數個步驟，首先會利用一個深度預測模型預測輸入之左圖的深度，取得深度圖後經過線性變換轉為視差圖，接著對視差圖做邊緣銳化，再將輸入之左圖和視差圖做 warping 和 filling，得到合成的右圖。而我們將整個流程結合我們提出的方法，並分為四個步驟：depth estimation、disparity process、image warping、background filling。

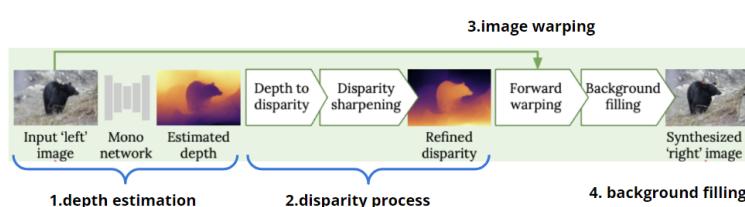


Figure 1: Reference pipeline

### 3 Methodology

### 3.1 Depth Estimation

在 Watson et al.[1] 的 work 中提出使用 MiDaS 預訓練模型 [2] 來為影像預測深度，而 MiDaS 模型又分為 Large、Hybrid 和 Small，模型越大，預測之深度越準確 (圖 2a)，但對於記憶體之需求相對較高也須較長之時間。受限於硬體之限制，我們使用 MiDaS Small 來為我們的影像進行深度之預測 (圖 2b)。

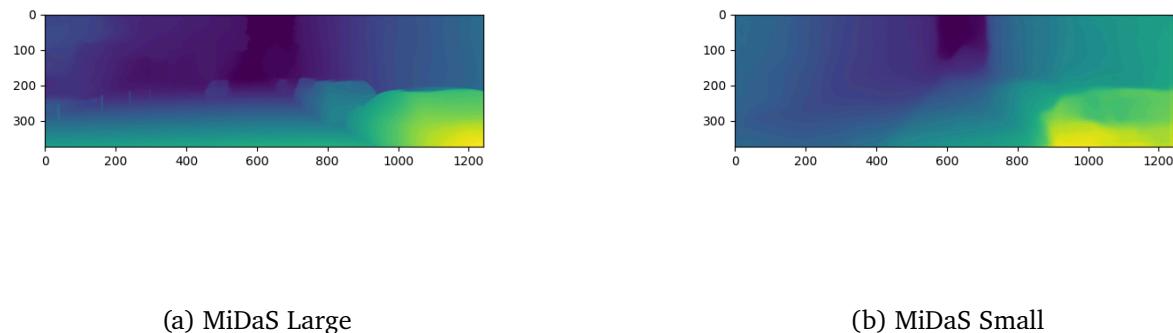


Figure 2: MiDaS

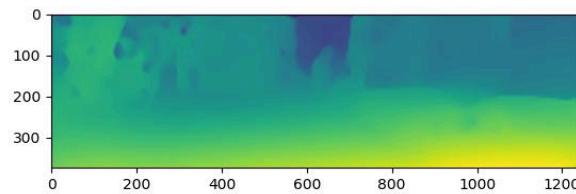


Figure 3: Disparity Estimation

進行完深度預測後，Wastont et al. 提出對深度進行線性轉換的做法以求得視差 (1)，然而此種做法中仰賴一個隨機變數  $s$  來進行計算，因此我們認為可以利用一個更好的模型來學習從不同影像之深轉換變為視差。參考 Liang et al.[3] 的 work 後，我們將預測之深度圖分成兩路進入下一階段之處理。在圖 4 中的上方為一個簡單的 Convolutional Layer，我們認為深度以及視差存在著強烈的反

關係，因此我們希望此 Convolutional Layer 可以學習將深度轉換為視差。而圖 4中的下方為一連串之 Convolution 處理，目的為讓 MiDaS Small 之深度預測結果能夠從模糊的影像（圖 2b）學到類似 MiDaS Large（圖 2a）中能夠捕捉物件之輪廓的能力。最後綜合上下兩路之結果，並產生如圖 3之結果。

$$\tilde{D} = \frac{sZ_{max}}{Z} \quad (1)$$

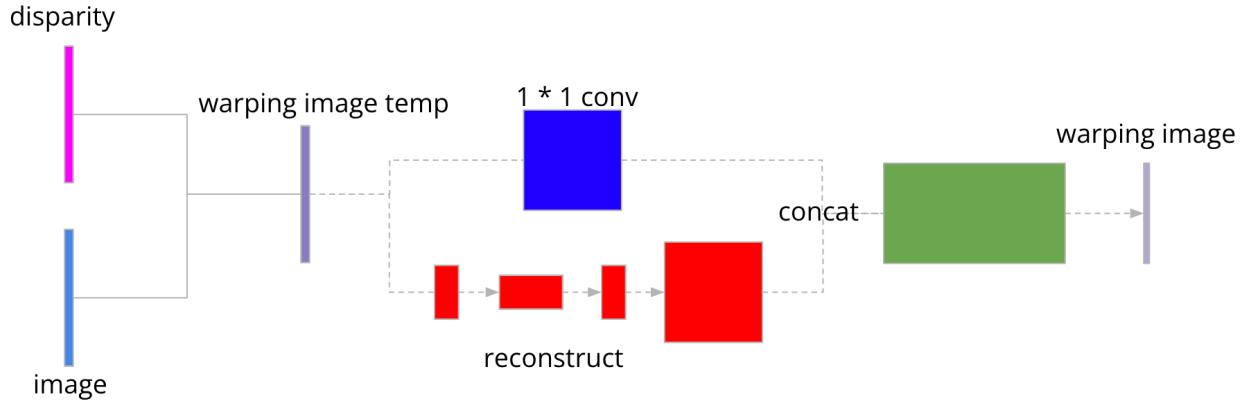


Figure 4: Model Architecture

### 3.2 Disparity Processing

在 3.1 中所提及之模型雖能預測視差，但其產生之視差仍有不完美之處。模型產生之視差圖對於遠近之視差不存在顯著差異，也就是說遠點與近點的視差相近，整體畫面類似平移（圖 5a、圖 5b）而非旋轉。因此我們提出新想法：

1. 將視差圖之值進行標準化（Normalization）
2. 將標準化之視差進行三次方的運算
3. 將三次方後之視差同乘原先視差圖中之最大值

進行 1 與 2 是為了對視差進行非線性轉換以拉開遠點與近點視差之距離。而 3 則是為了還原近點預測之視差，其結果如圖 6a 與 6b。



Figure 5: Before Disparity Processing



(a) Orginal Left Image



(b) Generated Right Image

Figure 6: After Disparity Processing

### 3.3 Warping and Background Filling

Warping 的部分相對簡單，我們採用 Bilinear Interpolation 來進行，然而此做法容易產生物件變形（如圖 7a 中之行人）之情況，而使得影像失真。而 Watson et al. 所提出之 Background Filling 作法 [1]，可以解決方才 Bilinear Interpolation 所存在的問題。但 Background Filling 做法是將圖片中存在遮擋（Occlusion）之像素利用影像集中任意影相來補齊，此作法對於序列（Sequential）影像和有前後影像存在劇烈晃動之情形並不適用，因為容易發生如圖 7b 中右下角處影像不夠平滑的問題。因此，我們做綜合上述兩種做法，將容易變形之處使用 Watson et al. 所提出之方法處理，而遮擋處則以 Bilinear Interpolation 來使該處影像足夠平滑，最後結果如圖 7c 所呈現。



(a) Pure Warping



(b) Background Filling



(c) Warping + Background Filling

Figure 7: Warping and Background Filling

## 4 Experiment & Analysis

我們將 KITTI Dataset 的左眼影像丟入我們的 pipeline 中生出右眼的雙目影像，並將此雙目影像當作雙目 SLAM 之輸入資料與使用 KITTI 影像之單目 SLAM 與雙目 SLAM 進行比較，其軌跡如圖 8 所示。我們將三種 SLAM 實作所預測之相機姿態與 ground truth 做誤差的計算，並將各自的 rotation 與 translation 誤差取中位數，並整理成表 1。

在圖 8 中可以看到，我們生成的雙目影像僅靠單目影像的訊息，就能有描繪相機軌跡的能力，雖仍有進步空間，但從表 1 中之誤差中位數可以得知我們的作法在 rotation 上能夠與使用 KITTI 之雙目影像的結果相近，甚至在 Seq 03 中超越了 KITTI 雙目影像的表現。

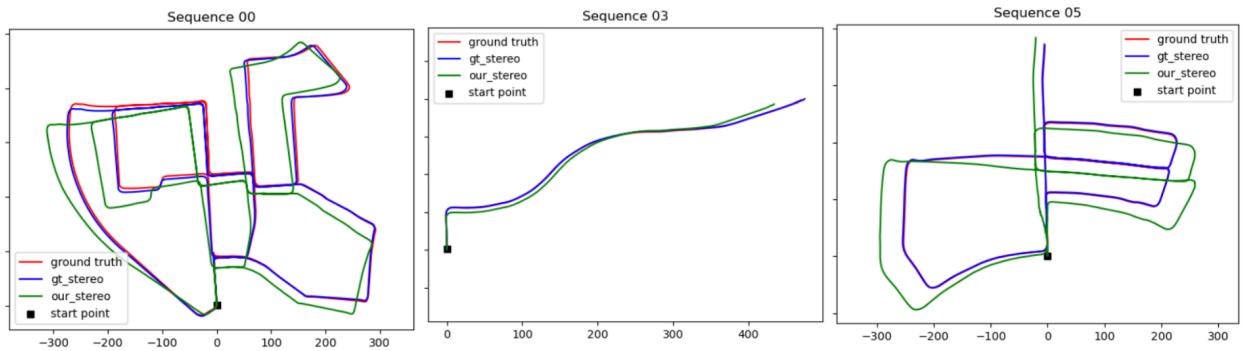


Figure 8: Estimated Trajectory

Table 1: Comparison of Different Implementation by Median Error

Sequence	Monocular		Our Stereo		Stereo	
	Rotation	Translation	Rotation	Translation	Rotation	Translation
00	1.67520	X	2.26992	23.60832	1.66759	6.20214
03	0.49867	X	<b>0.58842</b>	6.40959	0.71535	2.20344
05	1.02047	X	<b>0.89336</b>	20.96339	0.67948	1.48644

## 5 Conclusion

在簡單的路線上，我們的預測結果非常接近 ground truth，而複雜的路線大致上也都是路線偏移或是縮放差距，路徑的形狀與 ground truth 是一致的。我們的成果不僅能解決單一影像沒有深度、視差以及難估 t scale 的問題，甚至 rotation 能做的比上限還要好，因此可以得知先產生 disparity 再執行 SLAM 會比較好，也可以證實我們的 disparity 和 warping 的方式是有效的。

---

## 6 Discussion

雖然直接預測 disparity 會比較有效率，但先預測深度再預測 disparity 還是比較直觀，而之所以能僅靠一張圖片就能把 disparity 預測得那麼準確，也是有賴於 baseline 的限制，大部分的 baseline 和相機的內在參數都是固定的，就能依此優化，也能提升準確率。而根據觀察，好的 disparity 並不是像 ground truth 就好，還需要能幫助 warping 後的 image 更好取得 feature，所以加入 rule base 調整會比較好。

## References

- [1] J. Watson, O. M. Aodha, D. Turmukhambetov, G. J. Brostow, and M. Firman, “Learning stereo from single images,” in *European Conference on Computer Vision*. Springer, 2020, pp. 722–740.
- [2] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, “Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer,” *IEEE transactions on pattern analysis and machine intelligence*, 2020.
- [3] Z. Liang, Y. Feng, Y. Guo, H. Liu, W. Chen, L. Qiao, L. Zhou, and J. Zhang, “Learning for disparity estimation through feature constancy,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2811–2820.