Estymacja Głębi - WK Lab 9

Autor: Dariusz Max Adamski

Opis problemu

Problem komputerowej estymacji głębi polega na przypisaniu dla każdego piksela obrazu, wartości rzeczywistej opisującej jak daleko od obserwatora znajduje się punkt reprezentowany przez ten piksel. Jeden wariant tego problemu zakłada dane wejściowe w formie dwóch obrazów, pozyskanych przez dwa obiektywy do tworzenia zdjęć stereoskopowych, czyli w podobnej formie jaką posługują się ludzie przy widzeniu przestrzennym. Trudniejszy wariant problemu zakłada podanie tylko jednego obrazu jako dane wejściowe.

Wcześniejsze podejścia

Problem estymacji głębi dla pojedynczych obrazów nie był poruszany tak często jak estymacja głębi dla dwóch obrazów, a proponowane rozwiązania nie były tak dogłębne jak dla wariantu z obrazami stereo. Według autorów opisywanego artykułu wynika to z tego, że z dwóch obrazów z wystarczająco dokładnymi wzajemnymi odwzorowaniami można deterministycznie zbudować mapę głębi. Natomiast mapa głębi dla jednego obrazu może odpowiadać nieskończonej ilości rzeczywistych scen. Aby efektywnie znaleźć mapę głębi dla pojedynczego obrazu należy nie tylko brać pod uwagę kształty obiektów, perspektywę obrazu, ale też informacje kontekstowe, takie jak sens fizyczny sceny i oświetlenie. Pierwsze podejście do tego problemu wykorzystujące nadzorowane uczenie maszynowe zostało zaprezentowane w 2005 roku; wykorzystywało algorytm Markov Random Field (MRF) i osiągało subiektywnie dobre wyniki na zbiorze obrazów ze scenami leśnymi, miejskimi i w pomieszczeniach[1]. Wcześniejsze podejścia (np. artykuł z 1988 roku[2]) próbowały odwzorowywać głębię za pomocą informacji o oświetleniu. Nie osiągały one

[1] Saxena, Ashutosh, Sung H. Chung, and Andrew Y. Ng. "Learning depth from single monocular images." *NIPS*. Vol. 18. 2005.

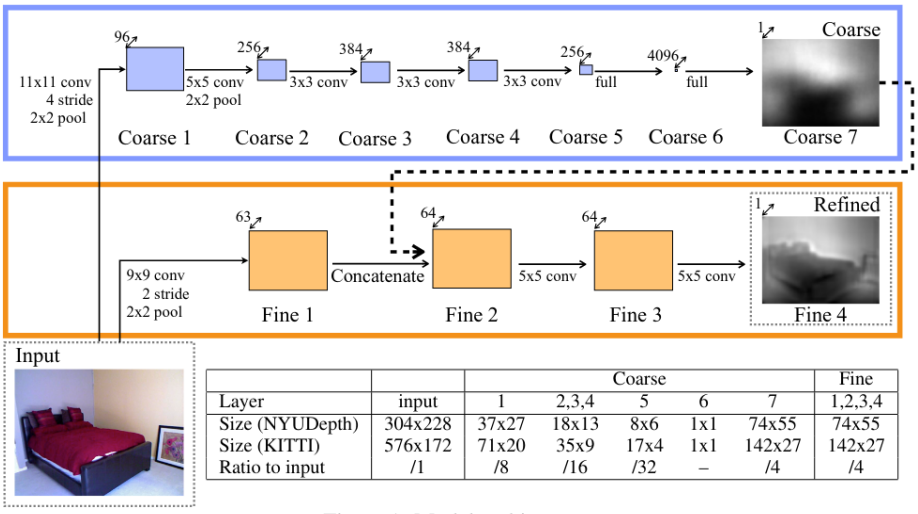
[2] Shao, Min, Tal Simchony, and Rama Chellappa. "New algorithms from reconstruction of a 3-d depth map from one or more images." *Proceedings CVPR'88: The Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 1988.

dobrych wyników przy nierównomiernym kolorze czy charakterystyce materiałów na obrazie.

Opis metody z wybranego artykułu

W artykule “Depth Map Prediction from a Single Image using a Multi-Scale Deep Network” (Eigen et al.; dostęp on-line[3]), jest opisane podejście do problemu estymacji głębi z pojedynczych obrazów RGB z wykorzystaniem nadzorowanego uczenia splotowych sieci neuronowych. Artykuł był opublikowany w 2014 roku na konferencji NIPS.

Według autorów, ważne do uzyskania dobrej informacji o głębi jest połączenie globalnych cech obrazu z lokalnymi cechami. Dlatego zamiast jednej sieci splotowej zastosowali dwie połączone. Jedna (niebieska na obrazku 1.) przewiduje mapę głębi o bardzo niskiej rozdzielczości, druga (pomarańczowa) przewiduje mapę głębi o wyższej rozdzielczości. Kluczowa w tej architekturze jest konkatenacja wyniku mapy głębi o niskiej rozdzielczości, w sieci o wysokiej rozdzielczości. W efekcie autorzy stworzyli architekturę *podobną* do architektury ResNet, która ukazała się rok później (2015, inni autorzy). Wszystkie warstwy ukryte korzystają z funkcji aktywacji ReLU, a ostatnia korzysta z aktywacji liniowej.



		Coarse					Fine
Layer	input	1	2,3,4	5	6	7	1,2,3,4
Size (NYUDepth)	304x228	37x27	18x13	8x6	1x1	74x55	74x55
Size (KITTI)	576x172	71x20	35x9	17x4	1x1	142x27	142x27
Ratio to input	/1	/8	/16	/32	–	/4	/4

Figure 1: Model architecture.

[3] https://papers.nips.cc/paper/2014/file/7bccfde7714a1ebadf06c5f4cea752c1-Paper.pdf

Kolejnym ważnym elementem proponowanego rozwiązania jest zastosowanie funkcji straty niezależnej od skali (scale-invariant mean squared error) do oceny modelu. Według autorów ważniejsze w estymacji głębi jest odwzorowanie wzajemnych względnych różnic w głębokości pomiędzy obszarami na obrazie (scale-invariant), niż ocenianie modelu na podstawie bezwzględnego błędu w ocenie głębi.

Dodatkowo, przy uczeniu modelu wykonywane są następujące augumentacje obrazków ze zbioru uczącego: zmiana skali, rotacja, translacja, zmiana intensywności koloru, odbicie lustrzane.

Osiągnięte wyniki

Autorzy do uczenia i oceny modelu wykorzystali dwa zbiory danych: NYU Depth v2 i KITTI. Dla zbioru NYU, w zbiorze treningowym znalazło się 120 tysięcy unikalnych obrazów (oversampling do 220K, żeby typy scen były zbalansowane), a w zbiorze testowym 694 obrazki. Dla zbioru KITTI, w zbiorze treningowym znalazło się 20 tysięcy unikalnych obrazów (oversampling do 40K). Autorzy nie wskazali ile obrazków znajdowało się w zbiorze testowym KITTI.

Uzyskany model autorzy porównali z algorytmem Make3D (Saxena et al.) wykorzystując miary RMSE (liniowa, logarytmiczna, log+scale-invariant), absolute relative difference i squared relative difference (wzory w artykule). Wartości metryk dla zbioru NYU Depth są zamieszczone w tabelce 1. Wizualizacje przykładów z obu zbiorów na następnej stronie.

	Mean	Make3D	Ladicky*&al*	Karsch*&al*	Coarse	Coarse + Fine	
threshold $\delta < 1.25$	0.418	0.447	0.542	–	**0.618**	0.611	higher
threshold $\delta < 1.25^2$	0.711	0.745	0.829	–	**0.891**	0.887	is
threshold $\delta < 1.25^3$	0.874	0.897	0.940	–	0.969	**0.971**	better
abs relative difference	0.408	0.349	–	0.350	0.228	**0.215**	
sqr relative difference	0.581	0.492	–	–	0.223	**0.212**	lower
RMSE (linear)	1.244	1.214	–	1.2	**0.871**	0.907	is
RMSE (log)	0.430	0.409	–	–	**0.283**	0.285	better
RMSE (log, scale inv.)	0.304	0.325	–	–	0.221	**0.219**	

Table 1: Comparison on the NYUDepth dataset

Autorzy uzyskali lepsze wyniki od Make3D i propozycji przedstawionych w dwóch innych artykułach. Co ciekawe, część sieci przewidująca mapę głębokości w niskiej rozdzielczości czasami osiąga trochę lepsze wyniki od części sieci łączącej cechy ogólne z lokalnymi. Nadal jednak obie części sieci uzyskują wyniki lepsze od rozwiązań innych autorów.

Subiektywnie: niektóre wyniki są dobre, chociaż rozdzielczość pozostawia wiele do życzenia. Czasami obiekty na pierwszym planie nie są zauważane. Czasami interpretacja głębi znacznie mija się z rzeczywistością.

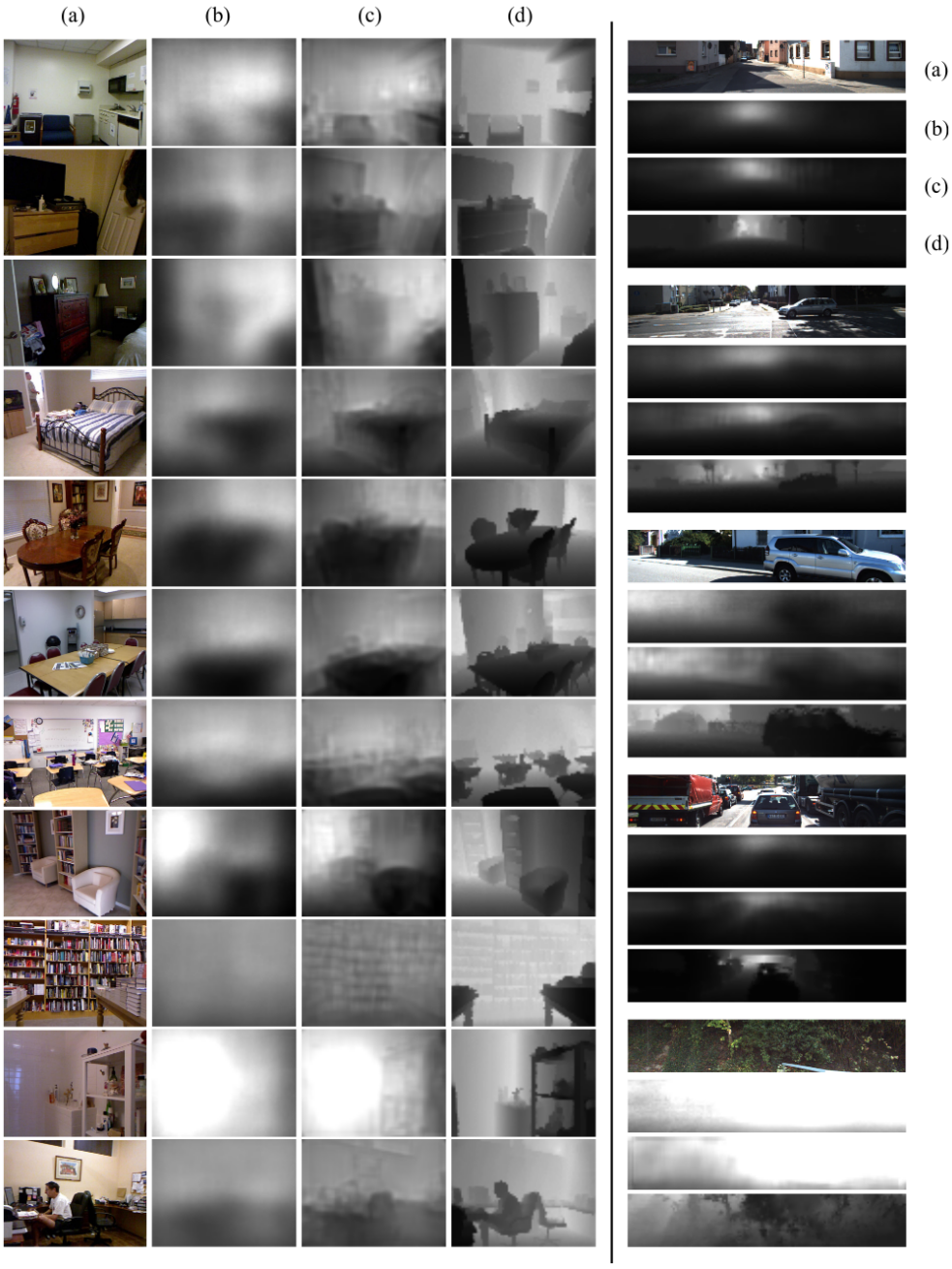


Figure 4: Example predictions from our algorithm. NYUDepth on left, KITTI on right. For each image, we show (a) input, (b) output of coarse network, (c) refined output of fine network, (d) ground truth. The fine scale network edits the coarse-scale input to better align with details such as object boundaries and wall edges. Examples are sorted from best (top) to worst (bottom).