

# ZPO – Odhad hloubkové mapy z dvojice snímků

Vít Tlustoš (xtlusty05), Jiří Vlasák (xvlasa15), Josef Kotoun (xkotou06)

4. května 2024

## 1 Zadání

### 1.1 Originální zadání

Odhadněte hloubkovou mapu z dvojice snímků. Snímky získejte například digitálním fotoaparátem ze dvou pozic vzájemně nepříliš vzdálených a pokud možno zaberte stejně zorné pole a obraz nerotujte, neměňte "zoom". Odhad hloubkové mapy provedte například tak (nebo úplně jinak, chcete-li) pomocí vzájemné korelace obrazů. Vždy malý výrez jednoho obrazu korelujte (například 16x16 pixelů) s okolím předpokládaného výskytu stejného motivu (části obrazu) ve druhém ze dvojice obrazů (například okolím 64x32 - 64 proto, že lze předpokládat posun motivu jen v horizontálním směru – viz též heslo "epipoláry v obrazu", 32 proto, že se přeče jen může něco vertikálně pohnout). Hloubkovou mapu lze odhadnout z rozdílnosti horizontálního posunu motivů na různých pozicích obrazu. Demonstrujte a diskutujte dosažené výsledky.

### 1.2 Upravené zadání

Naše řešení nabízí srovnání a přehled existujících metod a datasetů pro odhad hloubkových map. Jsou představeny: tradiční metody, založené na lokálním a semi-globálním přístupu hledání odpovídajících bodů – Block Matchin a Semi-Global Block Matching, metody založené na neuronových sítích – model HITNET pro odhad hloubkové mapy ze stereo snímků a model Depth Anything pro odhad hloubkové mapy z jednoho snímku. Následně je představen souhrn datasetů pro odhad hloubkové mapy s důrazem na datasety ETH3D, Middlebury 2014 a 2021. Také je diskutován způsob a princip akvizice anotací pro zmíněné datasety. Posledním výstupem je kvalitativní a kvantitativní analýza představených metod nad datasetem Middlebury 2014. Základní metoda Block Matching je implementována, ke srovnání ostatních metod jsou použity existující implementace.

Naším cílem je poskytnout přehled současných přístupů v této oblasti, včetně jejich klíčových charakteristik, výhod a potenciálních omezení. Naší ambicí je usnadnit výběr metody pro konkrétní aplikaci jak nám, tak i ostatním studentům. Důraz klademe hlavně na praktickou aplikovatelnost metod a datasetů.

## 2 Datasety (zpracoval Vít Tlustoš a Josef Kotoun)

V současné době existuje mnoho datasetů určených pro odhad hloubkové mapy. Tyto datasety lze charakterizovat na základě: rozsahu (jednotky/desítky snímků vs. statisíce snímků), metody pořízení hloubkové mapy (structured light, leaser, SLAM, jiné), počtu souvisejících snímků (monocular - jeden pohled do scény vs. stereo - dvojice částečně se překrývajících snímků vs. multi-view stereo - 3 a více částečně se překrývajících snímků) a doméně dat (syntetická data vs. data z reálného světa). Tabulka 1 obsahuje přehled nejběžněji používaných datasetů úlohy odhadu hloubkové mapy.

Dataset	Počet souv. snímků	Rozsah	Metoda pořízení hloubkové mapy	Doména
ETH3D	4	27 + 20 snímků, 5 + 5 videí, 13 + 12 scén	struktuované světlo	venkovní i vnitřní scény
Middlebury 2014	2	33 scén s několika snímkami	struktuované světlo	vnitřní scény
Middlebury 2021	2	24 scén s několika snímkami	struktuované světlo	vnitřní scény

Tabulka 1: Srovnání nejpoužívanějších datasetů pro úlohu odhadu hloubkové mapy.

Metoda pořízení hloubkové mapy přímo ovlivňuje přesnost hloubkové mapy. Zpravidla se používají následující metody:

- Triagulace stereo snímků – jsou pořízeny 2 či více částečně se překrývajících snímků scény a následně je za pomocí triangulace vypočtena hloubková mapa.
- Strukturované světlo – do scény je promítнут předem známý vzor (tečky, mřížka, pruh), zaznamenáván je objektem (scénou) deformovaný vzor.
- Měření laserem (LiDAR) - do scény je vyslán laserový paprsek a je měřen čas letu paprsku scénou.

V tabulce 2 lze vidět srovnání přístupů využívaných pro akvizici hloubkových map.

	<b>Strukturované světlo</b>	<b>Laserové měření (Lidar)</b>
Princip	Promítaný vzor	Doba letu
Měření	Analýza deformovaného vzoru	Měření času letu
Výhody	vysoká přesnost, vysoké rozlišení	dlouhý dosah, menší citlivost na světelné podmínky scény
Nevýhody	omezený dosah, citlivost na světelné podmínky scény	doba letu

Tabulka 2: Srovnání přístupů používaných pro pořízení hloubkové mapy.

Následující podkapitola 2.1 popisuje dataset Middlebury, který je využíván v rámci kvalitativních a kvantitativních experimentů v této práci.

## 2.1 Middlebury

Middlebury - 2021 Mobile stereo dataset se skládá z 24 různých vnitřních scén. Každá scéna byla zachycena pomocí Apple iPod touch 6G namontovaného na robotické paži, přičemž pro každou scénu jsou nabízeny dva snímky (two-view stereo) pořízené z různých pozic kamery. Hloubkové mapy jsou získány pomocí techniky strukturovaného osvětlení, které poskytuje přesné informace o hloubce pro každý pixel. Autoři dokonce tvrdí, že dosahují sub-pixel přesnosti. Nevýhodou této techniky je však, že není použitelná ve venkovním prostředí, proto tento dataset neobsahuje, žádné venkovní scény. Data pro každou scénu obsahují 2 snímky scény, hloubkovou mapu a kalibrační soubor. Middlebury - 2014 Stereo dataset je dataset pořízený stejnou technikou jako Middlebury - 2021 Mobile stereo dataset. Stejně tak jako jeho nástupce obsahuje pouze vnitřní scény. V tomto případě je nabízeno 33 scén.

## 2.2 ETH3D

Dataset ETH3D, vyvinutý v ETH Zurich, je určen pro srovnání algoritmů realizujících odhad hloubkových map. Nabízeny jsou testovací úlohy pro odhad hloubkové mapy z two-view stereo (2 snímky) a multi-view stereo (4 snímky). Výhodou tohoto datasetu je, že zachycuje, jak vnitřní, tak venkovní scény. Snímky scén jsou pořízeny více synchronizovaným kamerovým zařízením (typově DSLR) s různými zornými poli. Ke každé scéně jsou zachycena mračna bodů (point clouds) pomocí vysoce přesného laserového skeneru zachycující skutečnou 3D geometrii scén. Hloubková mapa je následně vypočtena z uvedených mračen bodů. Autoři aplikovali vícekrokový post-processing za účelem získání kvalitní hloubkové mapy. Lze konstatovat, že přesnost anotací je poměrně vysoká. Data pro oba typy testovacích úloh obsahují snímky, nabízené ve vysokém (RGB, 6000x4000) a nízkém (šedotónové, 940x490) rozlišení, kalibrační soubory, point clouds a hloubkové mapy.

## 3 Metody

Na základě stereo snímků lze pomocí níže představených metod odhadnout tzv. disparitu, což je hodnota reprezentující horizontální posun odpovídajících oblastí mezi dvěma snímkami. Z této hodnoty lze poté přímo odvodit hloubku pomocí triangulace. Výjimkou je metoda Depth Anything, která hloubkovou mapu odhaduje pouze nad jedním snímkem.

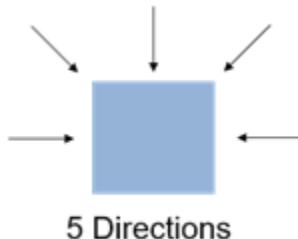
### 3.1 Block Matching (zpracoval Josef Kotoun)

Block matching je nejzákladnějším algoritmem pro odhad hloubkové mapy je tzv. block matching. Funguje na principu hledání shody mezi bloky pixelů ve dvou snímcích, které zachycují stejnou scénu s horizontalním posunem. Pro najítí shody je vypočítána podobnost v lokálním okolí druhého snímku, například pomocí korelace nebo sumy rozdílů (NCC - normalized cross correlation, SSD - Sum of Squared Difference, SAD - Sum of Absolute Difference).

Součástí odevzdaných zdrojových kódů je implementace metody Block Matching. Tato implementace má stejné rozhraní jako implementace z knihovny OpenCV – přijímá tedy parametry `maxDisparity` značící maximální hodnotu disparity a `blockSize` značící velikost okna, které je používáno při hledání odpovídající oblasti v druhém obrázku. Pro každý pixel je nejprve nalezena odpovídající poloha v druhém obrázku. Jako možné odpovídající pozice v druhém obrázku je bráno v potaz horizontální okolí o velikosti `maxDisparity`. Pro každou potenciální odpovídající polohu je spočítána suma rozdílů čtverců (SSD) v okolí, jehož velikost je dána parametrem `blockSize`. Jako odpovídající poloha oblasti je zvolena oblast s minimálním rozdílem a výsledná hodnota posunu (disparity) je vypočítána jako horizontální posun odpovídajících oblastí. Výsledná hodnota je poté naškálována do rozsahu 0-255 z původního rozsahu 0-`maxDisparity`. Výsledkem je mapa disparit o stejném rozlišení jako vstupní obrázek.

### 3.2 Semi-Global Block Matching (zpracoval Josef Kotoun)

Lokální přístup při hledání shody oblasti levého obrázku v pravém obrázku bere v potaz pouze lokální okolí potenciální shody. Lokální metody ale špatně detekují náhle změny hluboky, proto jsou preferovány globální přístupy, které využívají informace z celého obrázku. Semi-Global Block Matching kombinuje aspekty globálního a lokálního přístupu. Využívá sice okolních pixely, ale kombinuje informace z několika směrů. Pro experimenty je využita implementace SGBM z OpenCV, což je výpočetně efektivnější modifikace algoritmu SGM, který je originálně představen v publikaci [1]. V této implementaci je namísto hledání shody na úrovni pixelů hledána shoda na úrovni bloků a je bráno v potaz pouze pět směrů okolních pixelů, které jsou vizualizovány na obrázku 1. Primárními nastavitebnými parametry jsou velikost těchto bloků a rozsah disparit, ale nastavit lze i například počet směrů.



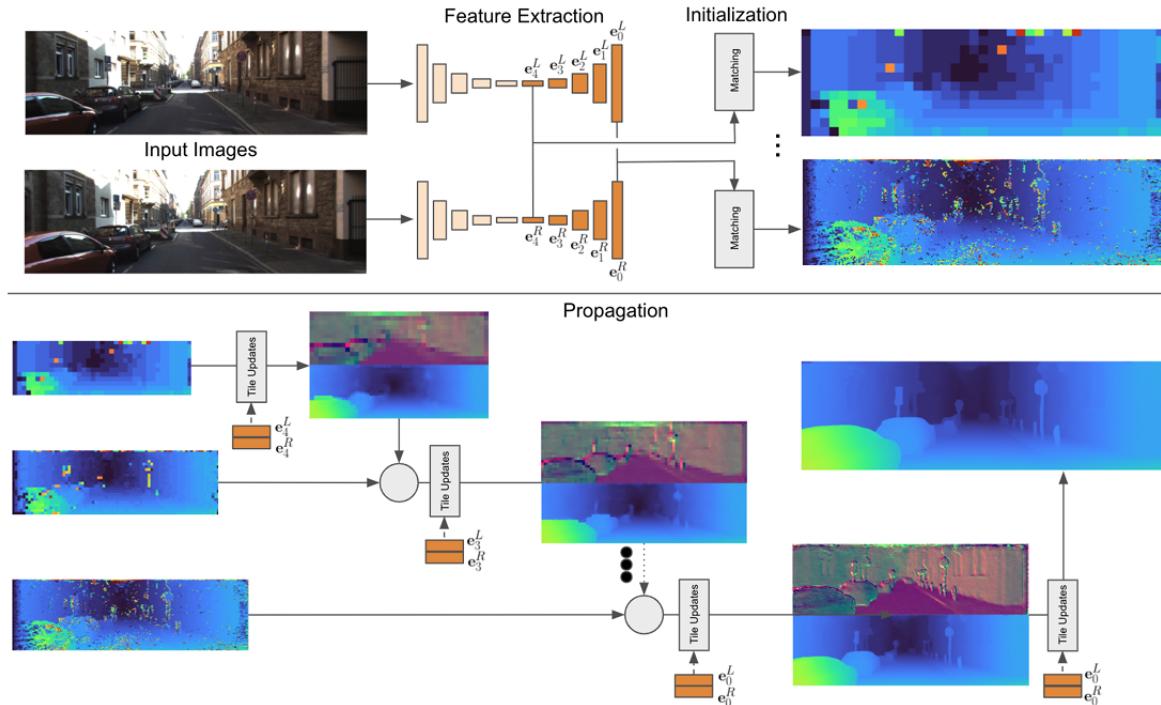
Obrázek 1: Směry použité v OpenCV implementaci SGBM

### 3.3 HITNet (zpracoval Jiří Vlasák)

HITNet [2] je hierarchická neuronová síť, která se používá pro odhad hloubky ze stereo snímků v reálném čase. Na rozdíl od mnoha předešlých přístupů založených na neuronových sítích, které, reprezentují scénu jako objem, a spoléhají se 3D konvoluce, HITNET explicitně nevytváří objem a místo toho se spoléhá na rychlý víceúrovňový inicializační krok, při kterém jsou získány informace z několika úrovní malé U-net sítě, a další kroky, při které nejsou součástí sítě a nejsou učené. Díky tomu má síť velmi málo parametrů a je mnohonásobně rychlejší než předchozí přístupy, při dosažení SOTA výsledků.

Architecture	Input Size	Number of Parameters	FLOPS
RAFT-Stereo	(512,512,3)	11.1162M	287.348G
FastACV-Plus	(512,512,3)	3.20329M	23.6084G
BGNet	(512,512,3)	5.31581M	23.5888G
PASMNet	(512,512,3)	7.81843M	-
PSMNet	(512,512,3)	5.22477M	292.052G
CREStereo	(512,512,3)	5.43295M	100.728G
HitNet	(512,512,3)	526.099K	14.4507G
GwcNet	(512,512,3)	6.51856M	244.779G

Obrázek 2: Srovnání různých neuronových sítí pro odhad hlubkové mapy ze stereo snímků. Převzato z [https://github.com/satya15july/depth\\_estimation\\_stereo\\_images](https://github.com/satya15july/depth_estimation_stereo_images)



Obrázek 3: Princip sítě HITNet, převzato z [2].

Navzdory nižší náročnosti a malému množství parametrů sítě, HITNet se při vydání v roce 2020 umístil na 1.-3. místě ve všech metrikách zveřejněných na webových stránkách ETH3D pro stereo přístupy, na 1. místě ve většině

metrik mezi všemi end-to-end učícími se přístupy na Middlebury-v3, na 1. místě v populárních benchmarcích KITTI 2012 a 2015 mezi zveřejněnými metodami rychlejšími než 100 ms.

Method	EPE px	Runtime s
HITNet XL	0.36	0.114
HITNet L	0.43	0.054
EdgeStereo [49]	0.74	0.32
LEAStereo [8]	0.78	0.3
GA-Net [61]	0.84	1.6
PSMNet [7]	1.09	0.41
StereoNet [25]	1.1	0.015

Obrázek 4: Srovnání výsledků sítě HITNet se state-of-the-art modely. Převzato z [2].

### 3.4 Depth Anything (zpracoval Vít Tlustoš)

Depth Anything [3] je nový (publikován 19.1.2024) open-source state-of-the-art monocular depth estimation model, za jehož vznikem stojí The University of Hong Kong, TikTok, Zhejiang Lab a Zhejiang University. Monocular depth estimation model znamená, že model na svém vstupu očekává pouze 1 snímek scény, ke kterému vypočte hloubkovou mapu. Tímto se model liší od tradičních způsobů pracujících na principu triangulace stereo snímků, metod založených na laserovém skenování či osvětlování scény strukturovaným světlem. Model tedy nevyžaduje žádný speciální hardware, jako je například LiDAR, IR projektor a IR kameru či dvojici kamer, což přispívá k jeho praktičnosti. Zároveň autoři pomocí podrobných experimentů prokázali, že jejich model dosahuje srovnatelných výsledků i v porovnání s two-view stereo modely/metodami, dále že jejich model je robustní (funguje dobře i na doménách na kterých nebyl učen). Dále autoři prokázali, že jejich model dobře generalizuje, tím, že udělili zero-shot evaluaci na následujících datasetech KITTI, NYUv2, ETH3D, DDAD a DIODE.

Architektura modelu se zakládá na encode-decoder Transformer architektuře. Konkrétně autoři použili DINOV2 model jako encoder a Dense Prediction Transformer (DPT) jako decoder.

Depth Anything model byl trénován na masivním datasetu obsahujícím 1.5M pseudo-anotovaných snímků. Anotace/hloubkové mapy pro anotované snímky byly získány většinou technikou triangulace stereo snímků, dále pak technikou structure from motion a LiDARem. Jednotlivé datasety jsou uvedeny v 5.

Dataset	Indoor	Outdoor	Label	# Images
Labeled Datasets				
BlendedMVS [76]	✓	✓	Stereo	115K
DIML [13]	✓	✓	Stereo	927K
HRWSI [67]	✓	✓	Stereo	20K
IRS [61]	✓		Stereo	103K
MegaDepth [33]		✓	SfM	128K
TartanAir [62]	✓	✓	Stereo	306K
Unlabeled Datasets				
BDD100K [81]		✓	None	8.2M
Google Landmarks [64]		✓	None	4.1M
ImageNet-21K [49]	✓	✓	None	13.1M
LSUN [80]	✓		None	9.8M
Objects365 [52]	✓	✓	None	1.7M
Open Images V7 [30]	✓	✓	None	7.8M
Places365 [87]	✓	✓	None	6.5M
SA-1B [27]	✓	✓	None	11.1M

Obrázek 5: Tabulka s přehledem datasetů, nad kterými byl trénován model Depth Anything

## Proces tréninku

Nejdříve byl natrénován teacher model. Jedná se o encoder-decoder model učený supervised-learning způsobem na zmíněných anotovaných datasetech. Jinými slovy pro každý RGB obrázek, byla k dispozici anotace – hloubková mapa. Použitou loss funkci lze vidět na obrázku 6.

$$\mathcal{L}_l = \frac{1}{HW} \sum_{i=1}^{HW} \rho(d_i^*, d_i), \quad \rho(d_i^*, d_i) = |\hat{d}_i^* - \hat{d}_i|$$

Obrázek 6: Loss funkce, která byla použita při tréninku teacher modelu

V principu se jedná mean absolute error, kde  $d_i^*$  je specificky zarovnaná predikovaná a  $d_i$  je skutečná (ground truth) hloubka pixelu i.

Následně autoři nechali teacher model vygenerovat hloubkové mapy tzv. pseudo labels pro všechny ostatní neanotované datasety. Tyto pseudo labels však obsahují stejné chyby teacher modelu, tedy student model by nebyl schopen předčít teacher model. Zde přichází na řadu klíčový poznatek tohoto paperu, který za pomocí specifických data augmentací a loss funkce umožňuje použití takto vygenerovaných pseudo labelů pro dotrénování základního (foundation) modelu:

- První trik je, že autoři aplikují silné modifikace na neaanotované snímky. Mezi tyto modifikace se patří: silné zkreslení barev (color distortion), color jittering, Gaussovské rozmazání a silné prostorové zkreslení (CutMix).
- Druhý trik je, že autoři k dříve zmíněné loss funkci, kterou v trochu modifikované variantě používají pro neanotované obrázky, přidali další člen Lfeat. V podstatě se jedná o penalizaci student modelu (fi je výstup encoderu student modelu) za to, že se odchylí od teacher modelu (fi je výstup encoderu zmraženého teacher modelu) na základě podobnosti jejich výstupu (cosine loss) ve feature space. Toto umravní způsob učení, takže je model schopen se naučit nové informace ze silně modifikovaných vstupů, ale není mu dovoleno příliš se odchýlit od teacher modelu a ztratit tak některé jeho schopnosti.

## 4 Metriky (zpracoval Jiří Vlasák)

**psm epe:** Průměrná absolutní chyba pixelů. Počítá pouze s pixely, která mají hodnotu nižší než hodnotu zvoleného prahu. V této práci se používá prah 192.

**bad X:** Procento pixelů s absolutní chybou větší než X.

## 5 Srovnání stereo metod (zpracovali Vít Tlustoš a Jiří Vlasák)

Jako základ pro kvantitativní porovnání výsledků jednotlivých metod jsme zvolili dataset Middlebury. Důvody výběru právě tohoto datasetu byly: dostatečný rozsah (15 snímků), vysoká přesnost anotací (sub-pixel) a skutečnost, že řada state-of-the-art metod/modelů je na něm testována, což umožňuje srovnat námi porovnávané metody s ostatními.

### 5.1 Augmentace

Součástí našeho řešení je také aplikace sady augmentací na vstupy (páry posunutých obrázků) jednotlivých testovaných metod a následné posouzení toho, jak daná augmentace ovlivnila výsledky metody. Tímto způsobem se snažíme zjistit robustnost metod vůči určitým typům augmentací. Augmentace jsme implementovali sami.

Následující kapitoly představují konkrétní implementace jednotlivých augmentací, zatímco v Tabulce 3 jsou uvedeny parametry/konfigurace těchto augmentací, jak jsou použity v kvantitativní analýze.

#### 5.1.1 Jas

$$I_{\text{nový}}(x, y) = \min(\max(0, I(x, y) + B), 255) \quad (1)$$

kde:

- $I_{\text{nový}}(x, y)$  představuje novou intenzitu pixelu na souřadnicích  $(x, y)$ .
- $I(x, y)$  je původní intenzita pixelu na souřadnicích  $(x, y)$ .
- B představuje hodnotu jasu přičtenou ke každému pixelu. Hodnota může být i záporná.

### 5.1.2 Kontrast

$$I_{\text{nový}}(x, y) = \min(\max(0, f \cdot (I(x, y) - 128) + 128), 255) \quad (2)$$

kde:

- $I_{\text{nový}}(x, y)$  představuje novou intenzitu pixelu na souřadnicích  $(x, y)$ .
- $I(x, y)$  je původní intenzita pixelu na souřadnicích  $(x, y)$ .
- $f$  reprezentuje koeficient kontrastu. Pro  $f > 1$  zvýší kontrast, jinak sníží.

### 5.1.3 Gaussovský Šum

$$I_{\text{nový}}(x, y) = I(x, y) + N(\mu, \sigma) \quad (3)$$

kde:

- $I_{\text{nový}}(x, y)$  představuje novou intenzitu pixelu na souřadnicích  $(x, y)$ .
- $I(x, y)$  je původní intenzita pixelu na souřadnicích  $(x, y)$ .
- $N(\mu, \sigma)$  reprezentuje normální rozdělení se střední hodnotou  $\mu$  a rozptylem  $\sigma$ .

### 5.1.4 Šum Sůl a Pepř

$$I_{\text{nový}}(x, y) = \begin{cases} I(x, y) & \text{pokud } U > p \\ 0 & \text{pokud } U \leq p \cdot r \\ 255 & \text{jinak} \end{cases} \quad (4)$$

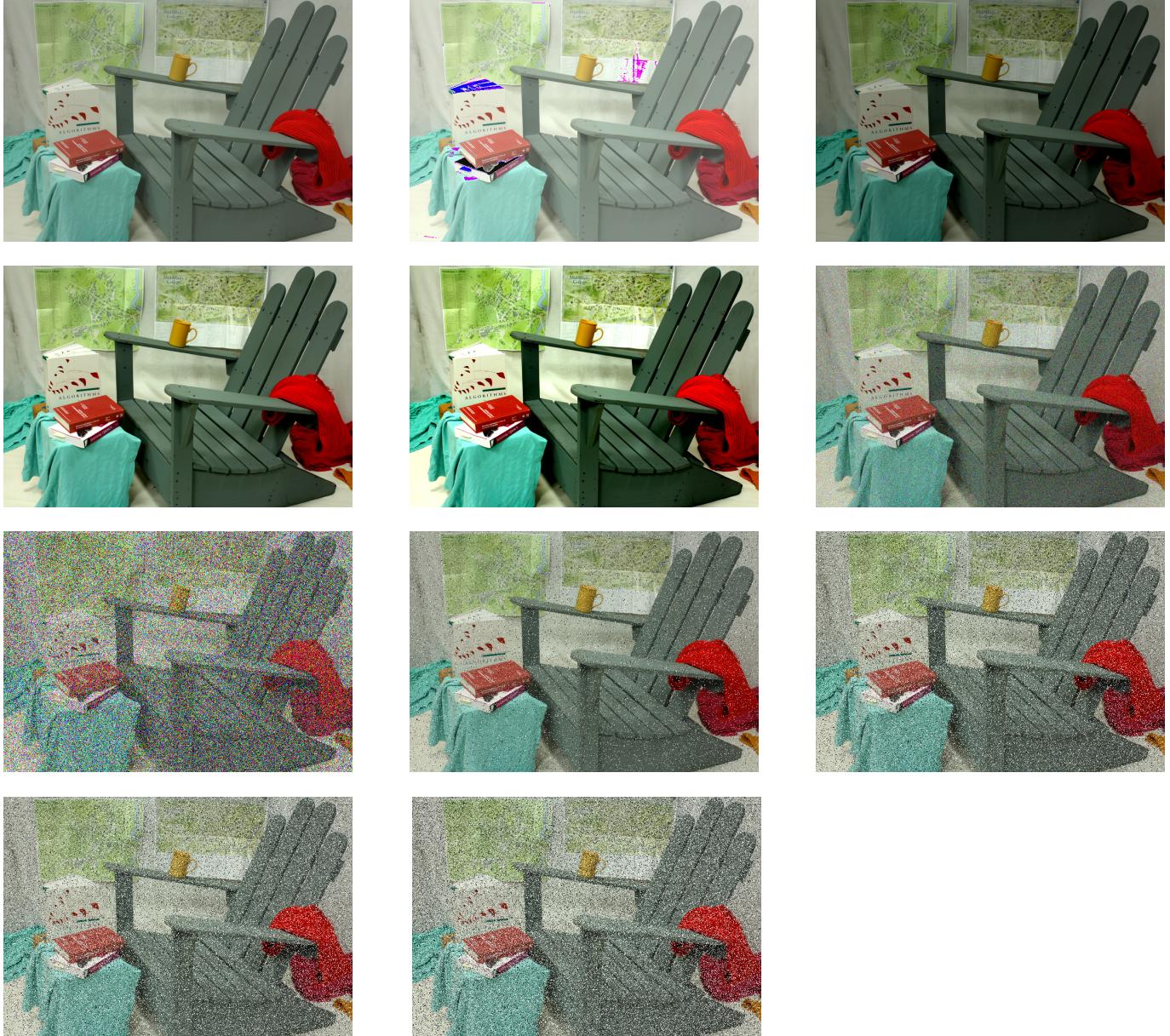
kde:

- $I_{\text{nový}}(x, y)$  představuje novou intenzitu pixelu na souřadnicích  $(x, y)$ .
- $I(x, y)$  je původní intenzita pixelu na souřadnicích  $(x, y)$ .
- $U$  je rovnoměrně rozdělená náhodná proměnná mezi 0 a 1, která určuje, zda je pixel ovlivněn šumem solí a pepře.
- $p$  je pravděpodobnost, že pixel bude vadný.
- $r$  je poměr pixelů ovlivněných solí (hodnota 255) a pepřem (hodnota 0).

### 5.1.5 Aplikované Augmentace

AUG	augmentace	parametry
0	nic	Neaplikuje žádnou augmentaci.
1	jas	$B = 50$ .
2	jas	$B = -50$ .
3	kontrast	$f = 1.5$
4	kontrast	$f = 2$
5	Gaussovský šum	$\mu = 0, \sigma = 0.1$
6	Gaussovský šum	$\mu = 0, \sigma = 0.5$
7	Sůl a Pepř šum	$p = 0.05, r = 0.5$
8	Sůl a Pepř šum	$p = 0.1, r = 0.5$
9	Sůl a Pepř šum	$p = 0.05, r = 0$
10	Sůl a Pepř šum	$p = 0.05, r = 1$

Tabulka 3: Použité augmentace a jejich parametry. AUG reprezentuje identifikátor augmentace.



Obrázek 7: Příklady argumentovaných obrázků s ID 0 (vlevo nahoře) až 10 (uprostřed dole).

## 5.2 Block Matching

Kvantitativně jsme prokázali, že vliv velikosti bloku a maximální uvažované disparity (parametry metody) mají významný vliv na dosahované výsledky. Toto tvrzení potvrzují data z Tabulky 4. Jako nejlepší se nám jeví velikost bloku 7 a maximální velikostí disparity 64.

B	D	<i>psm epe</i>	<i>bad</i> <sub>0.1</sub>	<i>bad</i> <sub>0.5</sub>	<i>bad</i> <sub>1.0</sub>	<i>bad</i> <sub>2.0</sub>	<i>bad</i> <sub>3.0</sub>
5	64	$18.10 \pm 12.13$	$99.68 \pm 0.11$	$97.65 \pm 0.76$	$78.68 \pm 8.71$	$53.74 \pm 14.53$	$52.39 \pm 14.62$
5	96	$18.74 \pm 8.24$	$99.69 \pm 0.11$	$97.78 \pm 0.69$	$79.86 \pm 6.78$	$56.21 \pm 9.79$	$54.99 \pm 9.75$
5	128	$20.20 \pm 6.96$	$99.70 \pm 0.11$	$97.88 \pm 0.64$	$81.29 \pm 5.82$	$59.50 \pm 8.58$	$58.37 \pm 8.53$
7	32	$26.21 \pm 13.76$	$99.69 \pm 0.21$	$98.06 \pm 1.18$	$85.80 \pm 8.09$	$70.60 \pm 13.54$	$69.47 \pm 13.80$
7	64	$17.45 \pm 13.08$	$99.71 \pm 0.12$	$97.82 \pm 0.79$	$76.93 \pm 9.82$	$49.37 \pm 15.62$	$48.18 \pm 15.73$
7	128	$18.35 \pm 7.37$	$99.71 \pm 0.13$	$97.90 \pm 0.75$	$79.03 \pm 6.92$	$54.15 \pm 9.60$	$53.09 \pm 9.57$
15	128	$17.90 \pm 8.34$	$99.64 \pm 0.21$	$97.49 \pm 1.19$	$78.06 \pm 7.27$	$52.92 \pm 10.31$	$51.65 \pm 10.39$

Tabulka 4: zobrazuje vliv parametrů metody Semi-Global Block Matching na výsledky bez aplikování žádné z augmentací. B reprezentuje velikost bloku, zatímco D maximální disparitu.

### 5.3 Semi-Global Block Matching

Kvantitativně jsme prokázali, že vliv velikosti bloku a maximální uvažované disparity (parametry metody) mají významný vliv na dosahované výsledky. Toto tvrzení potvrzuje data z Tabulky 5. Jako nejlepší se nám jeví velikost bloku 5 a maximální velikostí disparity 64.

B	D	<i>psm epe</i>	<i>bad</i> <sub>0.1</sub>	<i>bad</i> <sub>0.5</sub>	<i>bad</i> <sub>1.0</sub>	<i>bad</i> <sub>2.0</sub>	<i>bad</i> <sub>3.0</sub>
5	64	$11.93 \pm 10.79$	$99.85 \pm 0.09$	$99.18 \pm 0.36$	$98.03 \pm 0.74$	$67.77 \pm 10.36$	$35.04 \pm 17.11$
5	96	$11.61 \pm 6.69$	$99.85 \pm 0.09$	$99.21 \pm 0.40$	$98.09 \pm 0.84$	$68.41 \pm 7.64$	$36.08 \pm 11.15$
5	128	$12.40 \pm 5.24$	$99.85 \pm 0.09$	$99.19 \pm 0.45$	$98.02 \pm 0.98$	$69.61 \pm 6.35$	$39.06 \pm 10.00$
7	32	$20.82 \pm 13.41$	$99.72 \pm 0.18$	$98.56 \pm 0.97$	$96.98 \pm 2.02$	$80.04 \pm 9.88$	$61.66 \pm 15.07$
7	64	$12.10 \pm 11.00$	$99.85 \pm 0.07$	$99.18 \pm 0.34$	$98.01 \pm 0.72$	$68.16 \pm 10.16$	$35.44 \pm 16.87$
7	128	$12.48 \pm 5.40$	$99.84 \pm 0.11$	$99.16 \pm 0.50$	$97.98 \pm 1.06$	$69.91 \pm 6.29$	$39.41 \pm 9.84$
15	128	$13.07 \pm 5.80$	$99.82 \pm 0.12$	$99.03 \pm 0.55$	$97.60 \pm 1.27$	$70.91 \pm 6.08$	$42.24 \pm 9.51$

Tabulka 5: zobrazuje vliv parametrů metody Semi-Global Block Matching na výsledky bez aplikování žádné z augmentací. B reprezentuje velikost bloku, zatímco D maximální disparitu.

### 5.4 HitNet

Autoři HitNet nabízí několik natrénovaných checkpointů, včetně checkpointů trénovaných na Middlebury V3 datasetu, který používáme pro evaluaci. Protože byl tento checkpoint trénován na stejných obrázcích, nad kterými provádíme evaluaci, použili jsme místo toho checkpoint trénovaný na datasetu FlyingThings. Výsledky modelu HitNet jsou srovnány s ostatními metodami v následující kapitole.

### 5.5 Depth Anything

Model DepthAnything kvantitativně neporovnáváme, jelikož nezpracovává páry snímků, a tudíž na něj nelze aplikovat výše uváděné metriky zakladající se na disparitě.

### 5.6 Srovnání

V následujících srovnáních jsme použili následující parametry pro metodu Block Matching (BM):  $B = 7$  a  $D = 64$  a pro metodu Semi-Global Bloch Matching (SGBM) pak  $B = 5$  a  $D = 64$ .

#### 5.6.1 Žádná augmentace

Srovnání všech metod bez aplikace augmentací zobrazuje Tabulka 6. Je vidět, že HitNet překonává ostatní metody ve všech metrikách kromě  $bad_{3.0}$ . Je ale dobré podotknout, že jeho výsledky jsou mnohem méně konzistentní viz směrodatná odchylka ( $\pm$ ). Dále HitNet produkuje výrazně lepší výsledky než BM a SGBM už od pásma bad 0.1. Zároveň si lze všimnout, že metoda BM oproti SGBM produkuje lepší výsledky v pásmech bad 0.1 až 2.0.

Metoda	AUG	<i>psm epe</i>	<i>bad</i> <sub>0.1</sub>	<i>bad</i> <sub>0.5</sub>	<i>bad</i> <sub>1.0</sub>	<i>bad</i> <sub>2.0</sub>	<i>bad</i> <sub>3.0</sub>
BM	0	$17.45 \pm 13.08$	$99.71 \pm 0.12$	$97.82 \pm 0.79$	$76.93 \pm 9.82$	$49.37 \pm 15.62$	$48.18 \pm 15.73$
SGBM	0	$11.61 \pm 6.69$	$99.85 \pm 0.09$	$99.21 \pm 0.40$	$98.09 \pm 0.84$	$68.41 \pm 7.64$	$36.08 \pm 11.15$
HitNet	0	<b><math>8.65 \pm 10.54</math></b>	<b><math>87.90 \pm 7.25</math></b>	<b><math>61.67 \pm 20.67</math></b>	<b><math>51.39 \pm 24.09</math></b>	<b><math>42.09 \pm 26.24</math></b>	$36.83 \pm 27.17$

Tabulka 6: zobrazuje srovnání jednotlivých metod bez aplikování augmentací.

Obrázek 8 ukazuje výsledky jednotlivých metod na testovacím obrázku z Middlebury datasetu bez použití augmentací. Block Matching má problém s většími plochami a produkuje mnoho ”dér”. Metody SGBM a HitNet produkují značně lepší výsledky, přičemž HitNet má nejvyhlazenější predikce.



Obrázek 8: Výsledky metod BM (vlevo), SGBM (uprostřed) a HitNet (vpravo) na obrázku bez augmentací.

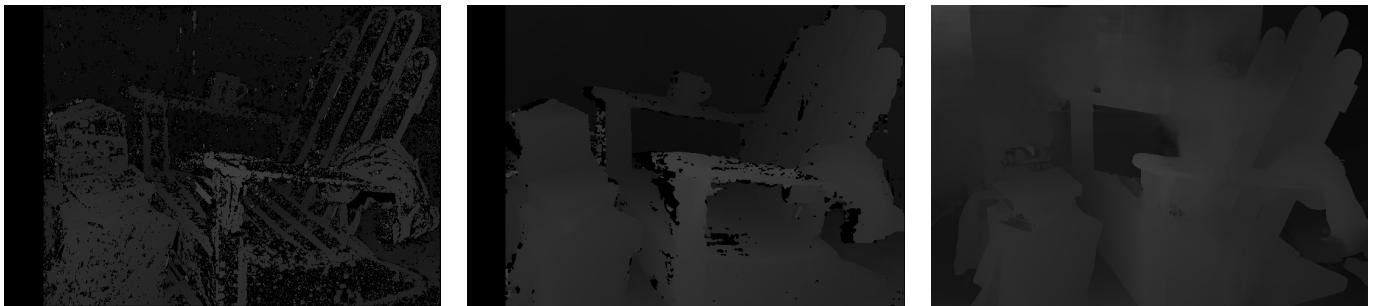
### 5.6.2 Jas

Výsledky jednotlivých metod po aplikování jasových augmentací jsou zaneseny v Tabulce 7. I zde vítězí metoda HitNet, zároveň si lze všimnout podobného chování (konzistentnost HitNet, ...) jako v případě 5.6.1.

Metoda	AUG	<i>psm epe</i>	<i>bad</i> <sub>0.1</sub>	<i>bad</i> <sub>0.5</sub>	<i>bad</i> <sub>1.0</sub>	<i>bad</i> <sub>2.0</sub>	<i>bad</i> <sub>3.0</sub>
BM	1	$18.68 \pm 13.02$	$99.68 \pm 0.14$	$97.70 \pm 0.86$	$78.75 \pm 8.89$	$53.44 \pm 14.81$	$52.14 \pm 14.96$
SGBM	1	$12.56 \pm 7.21$	$99.81 \pm 0.10$	$99.04 \pm 0.44$	$97.70 \pm 0.94$	$70.20 \pm 7.22$	$39.71 \pm 11.81$
HitNet	1	<b><math>11.29 \pm 11.75</math></b>	<b><math>90.88 \pm 6.49</math></b>	<b><math>70.66 \pm 17.40</math></b>	<b><math>59.66 \pm 20.84</math></b>	<b><math>48.81 \pm 23.65</math></b>	$42.55 \pm 25.20$
BM	2	$18.75 \pm 12.77$	$99.70 \pm 0.12$	$97.81 \pm 0.77$	$79.39 \pm 8.76$	$54.32 \pm 15.74$	$53.06 \pm 15.95$
SGBM	2	<b><math>12.73 \pm 8.20</math></b>	$99.83 \pm 0.09$	$99.12 \pm 0.40$	$97.86 \pm 0.84$	$70.24 \pm 7.55$	$40.00 \pm 13.35$
HitNet	2	$12.95 \pm 18.24$	<b><math>87.03 \pm 8.54</math></b>	<b><math>61.36 \pm 22.21</math></b>	<b><math>51.71 \pm 25.79</math></b>	<b><math>43.66 \pm 28.07</math></b>	<b><math>39.47 \pm 28.95</math></b>

Tabulka 7: zobrazuje srovnání jednotlivých metod po aplikování jasových augmentací.

Obrázek 9 ukazuje výsledky jednotlivých metod na testovacím obrázku z Middlebury datasetu se sníženým jasem o 50. Výsledky se u BM a SGBM příliš neliší od původního obrázku, u HitNet už ale výsledek není tak detailní a některé hrany ”splývají”.



Obrázek 9: Výsledky metod BM (vlevo), SGBM (uprostřed) a HitNet (vpravo) na obrázku s jasem sníženým o 50 (augmentace ID 2).

### 5.6.3 Kontrast

Výsledky jednotlivých metod po aplikování kontrastních augmentací jsou zaneseny v Tabulce 8. I zde vítězí metoda HitNet, zároveň si lze všimnout podobného chování (konzistentnost HitNet, ...) jako v případě 5.6.1.

Metoda	AUG	<i>psm epe</i>	<i>bad</i> <sub>0.1</sub>	<i>bad</i> <sub>0.5</sub>	<i>bad</i> <sub>1.0</sub>	<i>bad</i> <sub>2.0</sub>	<i>bad</i> <sub>3.0</sub>
BM	3	$18.81 \pm 13.14$	$99.69 \pm 0.13$	$97.76 \pm 0.82$	$79.26 \pm 8.87$	$54.04 \pm 15.86$	$52.79 \pm 16.08$
SGBM	3	$13.16 \pm 8.31$	$99.81 \pm 0.14$	$99.02 \pm 0.63$	$97.62 \pm 1.24$	$70.48 \pm 7.86$	<b>40.76 ± 13.58</b>
HitNet	3	<b>11.42 ± 13.93</b>	<b>88.47 ± 8.64</b>	<b>65.42 ± 20.94</b>	<b>55.59 ± 24.49</b>	<b>46.87 ± 26.56</b>	$41.55 \pm 27.71$
BM	4	$20.34 \pm 13.02$	$99.67 \pm 0.13$	$97.73 \pm 0.81$	$81.26 \pm 8.03$	$58.96 \pm 15.00$	$57.67 \pm 15.23$
SGBM	4	<b>14.47 ± 8.63</b>	$99.80 \pm 0.09$	$98.95 \pm 0.44$	$97.48 \pm 0.95$	$72.58 \pm 7.79$	<b>45.75 ± 14.36</b>
HitNet	4	$14.49 \pm 17.49$	<b>90.64 ± 7.02</b>	<b>70.57 ± 18.70</b>	<b>61.26 ± 22.73</b>	<b>52.75 ± 25.71</b>	$47.71 \pm 27.28$

Tabulka 8: zobrazuje srovnání jednotlivých metod po aplikování kontrastních augmentací.

Obrázek 10 ukazuje výsledky jednotlivých metod na testovacím obrázku z Middlebury datasetu s dvojnásobným kontrastem. Výsledky se u BM a SGBM příliš neliší od původního obrázku, u HitNet už ale začínají vznikat "díry" v židli a začíná se projevovat jeho vyšší závislost na původním datasetu.



Obrázek 10: Výsledky metod BM (vlevo), SGBM (uprostřed) a HitNet (vpravo) na obrázku s dvojnásobným kontrastem (augmentace ID 4).

### 5.6.4 Gaussovský Šum

Výsledky jednotlivých metod po aplikování Gaussovského šumu augmentací jsou zaneseny v Tabulce 9. S tímto šumem si nejlépe poradila metoda SGBM. Zajímavé je, že metoda HitNet v tomto případě úplně selhala, což dokazuje její náchylnost k Gaussovskému šumu.

Metoda	AUG	<i>psm epe</i>	<i>bad</i> <sub>0.1</sub>	<i>bad</i> <sub>0.5</sub>	<i>bad</i> <sub>1.0</sub>	<i>bad</i> <sub>2.0</sub>	<i>bad</i> <sub>3.0</sub>
BM	5	$30.88 \pm 10.74$	$99.82 \pm 0.07$	$98.87 \pm 0.46$	$96.35 \pm 1.85$	$92.91 \pm 3.45$	$92.04 \pm 3.62$
SGBM	5	<b>22.87 ± 7.83</b>	<b>99.68 ± 0.09</b>	<b>98.36 ± 0.45</b>	<b>96.32 ± 0.94</b>	<b>86.38 ± 3.91</b>	<b>77.37 ± 6.76</b>
HitNet	5	$94.00 \pm 14.33$	$99.99 \pm 0.02$	$99.96 \pm 0.08$	$99.91 \pm 0.16$	$99.82 \pm 0.34$	$99.71 \pm 0.57$
BM	6	$32.26 \pm 10.44$	$99.92 \pm 0.02$	$99.58 \pm 0.10$	$99.15 \pm 0.21$	$98.34 \pm 0.41$	$97.64 \pm 0.52$
SGBM	6	<b>30.92 ± 9.56</b>	<b>99.89 ± 0.02</b>	<b>99.41 ± 0.11</b>	<b>98.78 ± 0.26</b>	<b>97.30 ± 0.64</b>	<b>95.91 ± 1.00</b>
HitNet	6	$143.08 \pm 12.76$	$100.00 \pm 0.00$	$100.00 \pm 0.00$	$100.00 \pm 0.00$	$100.00 \pm 0.00$	$100.00 \pm 0.01$

Tabulka 9: zobrazuje srovnání jednotlivých metod po aplikování Gaussovského šumu.

Obrázek 11 ukazuje výsledky jednotlivých metod na testovacím obrázku z Middlebury datasetu s přidaným gaussovským šumem s  $\mu = 0$ ,  $\sigma = 0.1$ . Výsledky jsou u všech metod už těžko rozpoznatelné, přičemž u SGBM lze ztěží rozpoznat pouze výrazné hrany z původního obrázku. Nejzajímavější posun je vidět u HitNet, který pravděpodobně vůbec nerozpoznává vstupní obrázky a proto pouze interpoluje průměrnou hodnotu snímku.

### 5.6.5 Sůl a Pepř

Výsledky jednotlivých metod po aplikování šumu typu Sůl a Pepř jsou zaneseny v Tabulce 10. S tímto šumem si nejlépe poradila metoda SGBM. V případě tohoto typu šumů metod HitNet vůbec nefunguje, což v kombinaci s náchylností ke



Obrázek 11: Výsledky metod BM (vlevo), SGBM (uprostřed) a HitNet (vpravo) na obrázku s přidaným gaussovským šumem s  $\mu = 0$ ,  $\sigma = 0.1$  (augmentace ID 5).

Gaussovskému šumu ukazuje, že tato metoda není robustní vůči šumu.

Metoda	AUG	<i>psm epe</i>	<i>bad</i> <sub>0.1</sub>	<i>bad</i> <sub>0.5</sub>	<i>bad</i> <sub>1.0</sub>	<i>bad</i> <sub>2.0</sub>	<i>bad</i> <sub>3.0</sub>
BM	7	$27.36 \pm 11.11$	$99.70 \pm 0.11$	<b>98.00</b> $\pm 0.73$	<b>91.39</b> $\pm 3.30$	$82.55 \pm 6.07$	$81.28 \pm 6.25$
SGBM	7	<b>20.57</b> $\pm 7.53$	<b>99.67</b> $\pm 0.12$	$98.23 \pm 0.61$	$96.06 \pm 1.30$	<b>82.38</b> $\pm 5.08$	<b>69.51</b> $\pm 8.65$
HitNet	7	$162.87 \pm 17.23$	$100.00 \pm 0.00$	$99.99 \pm 0.01$	$99.99 \pm 0.01$	$99.98 \pm 0.02$	$99.96 \pm 0.04$
BM	8	$31.85 \pm 10.65$	$99.85 \pm 0.06$	$99.15 \pm 0.29$	$97.95 \pm 0.78$	$95.99 \pm 1.49$	$95.14 \pm 1.67$
SGBM	8	<b>27.08</b> $\pm 8.49$	<b>99.75</b> $\pm 0.09$	<b>98.69</b> $\pm 0.44$	<b>97.26</b> $\pm 0.91$	<b>92.80</b> $\pm 2.49$	<b>88.70</b> $\pm 4.01$
HitNet	8	$178.82 \pm 15.48$	$100.00 \pm 0.00$				
BM	9	$32.16 \pm 10.60$	$99.87 \pm 0.05$	$99.32 \pm 0.23$	$98.44 \pm 0.58$	$96.96 \pm 1.09$	$96.18 \pm 1.26$
SGBM	9	<b>28.07</b> $\pm 8.64$	<b>99.78</b> $\pm 0.08$	<b>98.86</b> $\pm 0.40$	<b>97.60</b> $\pm 0.80$	<b>94.07</b> $\pm 2.01$	<b>90.85</b> $\pm 3.22$
HitNet	9	$179.20 \pm 15.38$	$100.00 \pm 0.00$				
BM	10	$32.40 \pm 10.55$	$99.90 \pm 0.03$	$99.49 \pm 0.16$	$98.89 \pm 0.37$	$97.87 \pm 0.66$	$97.16 \pm 0.80$
SGBM	10	<b>29.45</b> $\pm 9.05$	<b>99.82</b> $\pm 0.06$	<b>99.09</b> $\pm 0.28$	<b>98.16</b> $\pm 0.56$	<b>95.77</b> $\pm 1.36$	<b>93.54</b> $\pm 2.13$
HitNet	10	$180.27 \pm 14.86$	$100.00 \pm 0.00$				

Tabulka 10: zobrazuje srovnání jednotlivých metod po aplikování šumu typ Sůl a Pepř.

Obrázek 12 ukazuje výsledky jednotlivých metod na testovacím obrázku z Middlebury datasetu s přidaným šumem typu pepř a sůl s  $p = 0.05$ ,  $r = 0.5$ . Výsledky jsou u všech metod podobné jako u gaussovského šumu, tedy už těžko rozpoznatelné, přičemž u SGBM lze ztěží rozpoznat pouze výrázné hrany z původního obrázku. HitNet stejně jako u gaussovského šumu pouze interpoluje průměrnou hodnotu snímku.



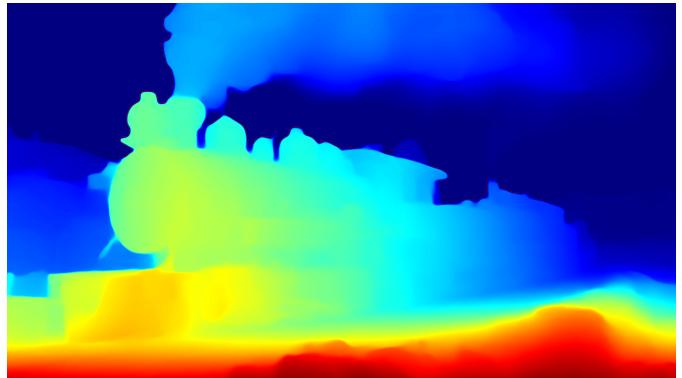
Obrázek 12: Výsledky metod BM (vlevo), SGBM (uprostřed) a HitNet (vpravo) na obrázku s přidaným šumem typu pepř a sůl s  $p = 0.05$ ,  $r = 0.5$  (augmentace ID 7).

## 6 Kvalitativní analýza výsledků Depth Anything (zpracoval Vít Tlustoš)

Abychom kvalitativně zhodnotili model DepthAnything, nechali jsme jej zpracovat následující 3 obrázky.

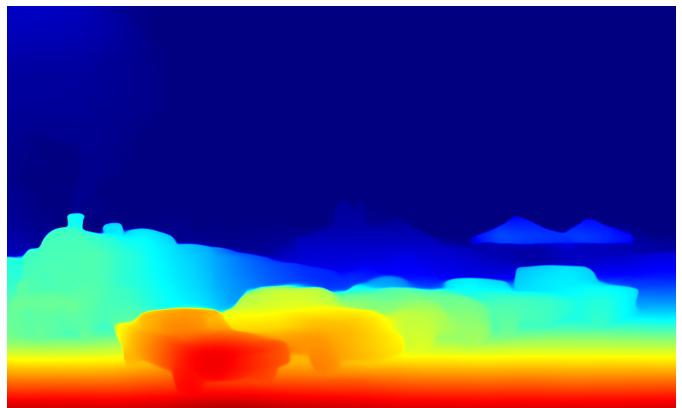
## 6.1 Běžný Snímek

Příklad níže ukazuje, že model je schopen přesně stanovit hloubku z jednoho snímku. Model korektně určuje i hloubku okolních objektů jako jsou domy.



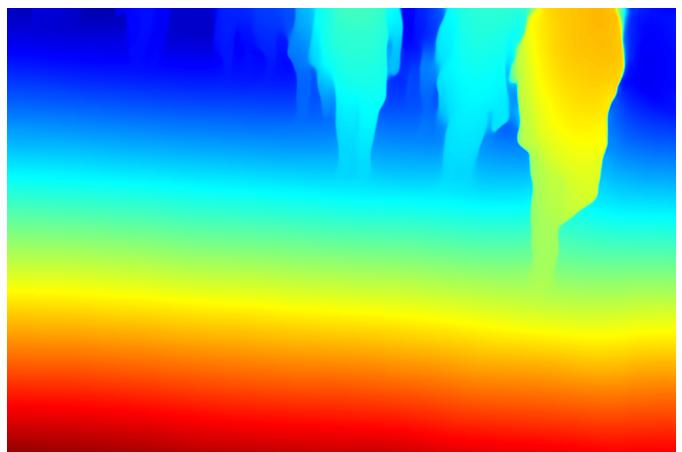
## 6.2 Kreslený Snímek

Příklad níže ukazuje, že model je schopen přesně stanovit hloubku z jednoho snímku, a to dokonce i v případě kresleného snímku. Podoktněme, že model korektně určil hloubku pro vzdálené objekty jako je hrad či stánky v pozadí.



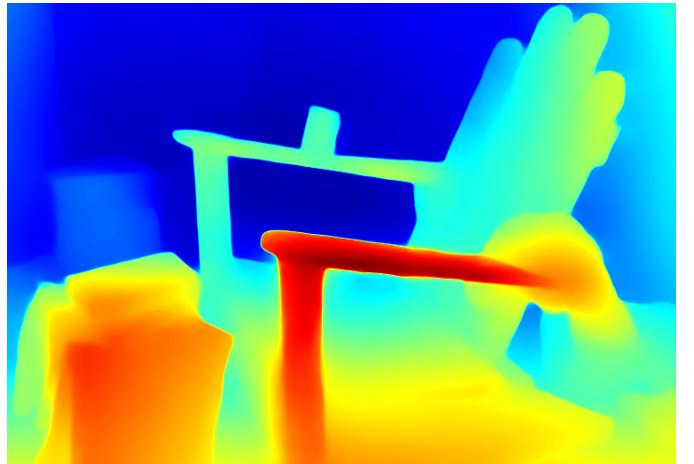
## 6.3 Stíny

Příklad níže ukazuje, jak se model vypořádal se stíny. Model korektně určil hloubku pouze pro reálné obejky (lidi) a stíny odfiltroval.



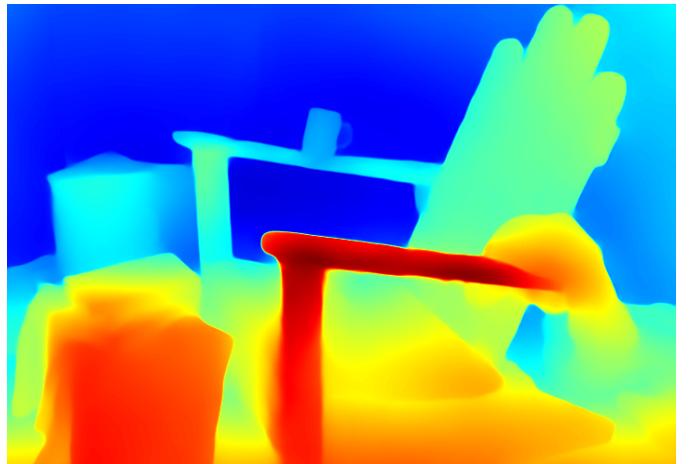
## 6.4 Gaussovský Šum

Příklad níže ukazuje, jak se model vypořádal se Gaussovským šumem ( $\mu = 0$  a  $\sigma = 0.1$ ). Jako jediný z testovaných přístupů, model DepthAnything dosáhl očekávaného výsledku, což ilustruje jeho odolnost vůči tomuto typu šumu.



## 6.5 Sůl a Pepř

Příklad níže ukazuje, jak se model vypořádal se šumem typu sůl a pepř ( $p = 0.05$  a  $r = 0.5$ ). I v tomto případě jako jediný z testovaných přístupů, model DepthAnything dosáhl očekávaného výsledku, což ilustruje jeho odolnost vůči tomuto typu šumu.



## 6.6 Zhodnocení

Model DepthAnything prokázal robustní chování napříč doménami, odolnost vůči různým formám šumu a správné chování v případě stínů. Naměřené výsledky u této metody bohužel nelze přímo porovnat s ostatními metodami, jelikož metoda pracuje pouze s jediným snímkem. Avšak kvalitativní posouzení nasvědčuje výborným výsledkům. Pokud bychom implementovali odhad hloubkové mapy, pravděpodobně bychom využili tuto metodu.

## Reference

- [1] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007.
- [2] V. Tankovich, C. Häne, Y. Zhang, A. Kowdle, S. Fanello, and S. Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14357–14367, 2021.
- [3] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.