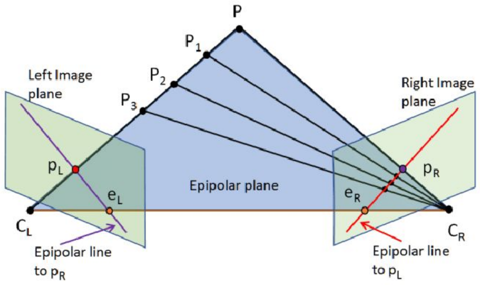
Research Progress:

Supervised Monocular Depth Estimation via Stacked Generalization

School of Environment and Society
Department of Transdisciplinary Science and Engineering
Yamashita Laboratory
CHINCHUTHAKUN WORAMETH

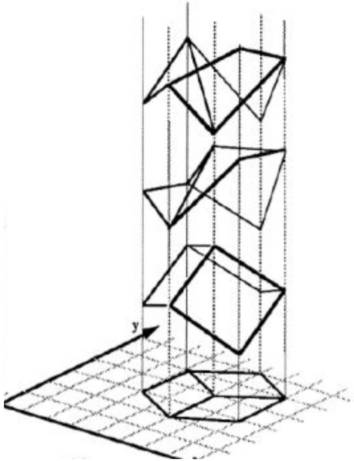
Background: Monocular Depth Estimation

- **Active** → depth sensors based on wave reflection
- **Passive** → use images from different perspectives to predict depth map
  - Stereo (2), Multiview (2+) → Near-perfect Approximation via epipolar geometry
  - Monocular (1) → Ill-posed problem
- Can provide a cost, space, and energy efficient alternative
  - Extremely useful in small robotic platforms



Epipolar geometry [1]

RGB image [2] and Depth map [3]



Why it's hard?[4]

[1] D. Chotrov, Z. Uzunova, Y. Yordanov, and S. Maleshkov, "Mixed-reality spatial configuration with azed mini stereoscopic camera," 2018. Available: https://www.researchgate.net/publication/329443348_Mixed-Reality_Spatial_Configuration_with_a_ZED_Mini_Stereoscopic_Camera
[2] N. Silberman, D. Hoiem, P. Kohli, and R.Fergus, "Indoor segmentation and support inference from rgbd images," in *Computer Vision – ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, andC. Schmid, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 746–760.
[3] C. Chaijirawiwat, "Monocular Depth Estimation via Transfer Learning and Multi-Task Learning with Semantic Segmentation," Bachelor's thesis, Tokyo Institute of Technology, Tokyo, Jul. 2019.
[4] D. Tan, "Depth estimation: Basics and Intuition," Medium, 12-Feb-2021. [Online]. Available: https://towardsdatascience.com/depth-estimation-1-basics-and-intuition-86f2c9538cd1. [Accessed: 14-Oct-2021].

Background: Monocular Depth Estimation (2)

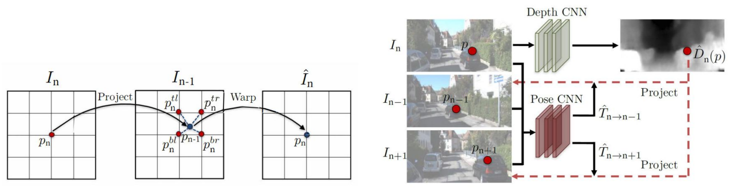
- How can we perceive depth in this painting?

Gustave Caillebotte's painting of a rainy street in Paris [1]

[1] "Artists and depth perception," *Psychology Today*. [Online]. Available: https://www.psychologytoday.com/gb/blog/ulterior-motives/201104/artists-and-depth-perception. [Accessed: 08-Nov-2021].

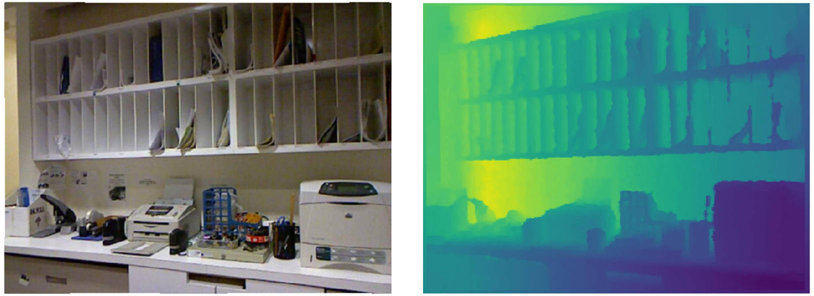
Background: Monocular Depth Estimation (3)

- We perceive depth subconsciously → difficult to mathematically describe how
- Deep learning approaches
  - Supervised → Use ground truth depth maps
  - Unsupervised → Use geometric constraints between frames in a monocular videos
  - Semi-supervised → Use stereo image pairs

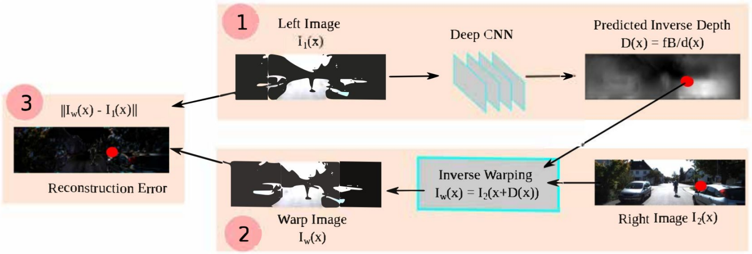


(a) An illustration of image warping process; (b) An illustration of unsupervised monocular depth.

Unsupervised problem setting [3]



Supervised problem setting [1, 2]

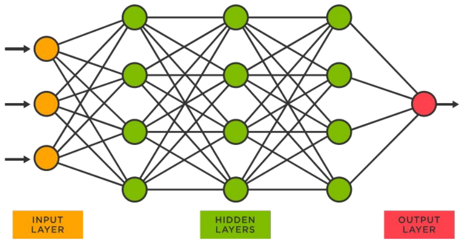


Semi-supervised problem setting [3]

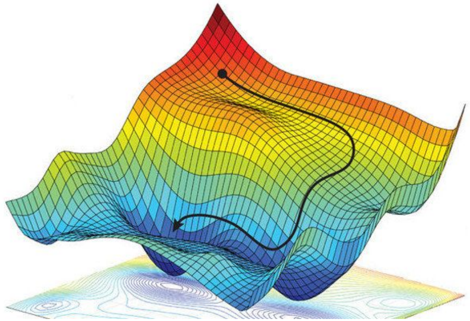
[1] N. Silberman, D. Hoiem, P. Kohli, and R.Fergus, "Indoor segmentation and support inference from rgbd images," in *Computer Vision – ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, andC. Schmid, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 746–760.
[2] C. Chaijirawiwat, "Monocular Depth Estimation via Transfer Learning and Multi-Task Learning with Semantic Segmentation," Bachelor's thesis, Tokyo Institute of Technology, Tokyo, Jul. 2019.
[3] C. Zhao, Q. Sun, C. Zhang, Y. Tang, and F. Qian, "Monocular depth estimation based on deep learning: An overview," *CoRR*, vol. abs/2003.06620, 2020. [Online]. Available: https://arxiv.org/abs/2003.0662017

Background: Deep Neural Network (DNN)

- **ML model** learns parameter θ to approximate $f(y|\theta) = x$
- **Neural network (NN)** is a specific type of ML models
  - Logistic regression is basically a one-layer neural network
- **Deep NN (DNN)** is NN with more than one layers
  - We often use **Convolutional Neural Network (CNN)** to process images since it can capture (local) spatial information
- Train by minimizing a **loss function** using variations of **gradient descend**
- **Transfer learning** (reuse NN's parameters in similar tasks)
  - **Freezing** → Completely reuse
  - **Fine-tuning** → Reuse with slight adjustments
  - We call NN being transferred as **pretrained NN**



Deep Neural Network [1]

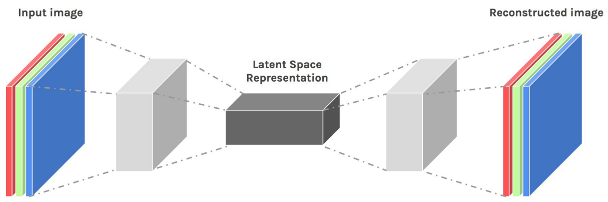


Gradient descend [2]

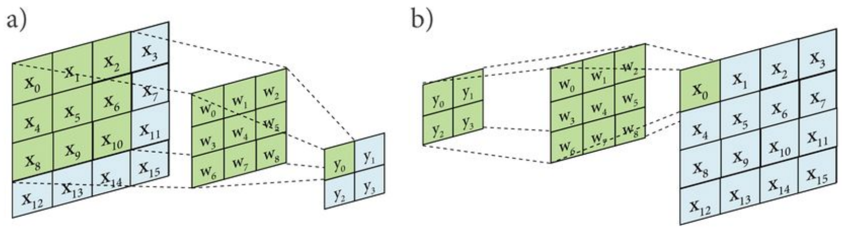
[1] "What is a neural network?," *TIBCO Software*. [Online]. Available: https://www.tibco.com/reference-center/what-is-a-neural-network. [Accessed: 14-Oct-2021].
[2] A. Amini, A. Soleimany, S. Karaman, and D. Rus, "Spatial uncertainty sampling for end-to-end control," *CoRR*, vol. abs/1805.04829, 2018. [Online]. Available: http://arxiv.org/abs/1805.04829

Background: Encoder-Decoder Framework

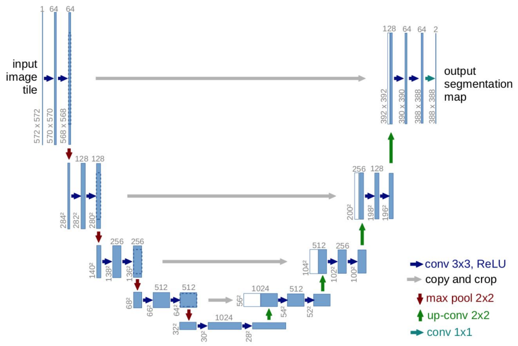
- CNN keeps reducing input dimension, but we need depth map to have the same size with image
- Just append a CNN (**Encoder**) and an inverted CNN (**Decoder**) together!
  - In practice, we use **interpolation + convolution** instead
  - **U-Net**, which serves as a baseline, also employs **residual connection**



Encoder-Decoder [2]



Convolution and Transposed convolution [1]

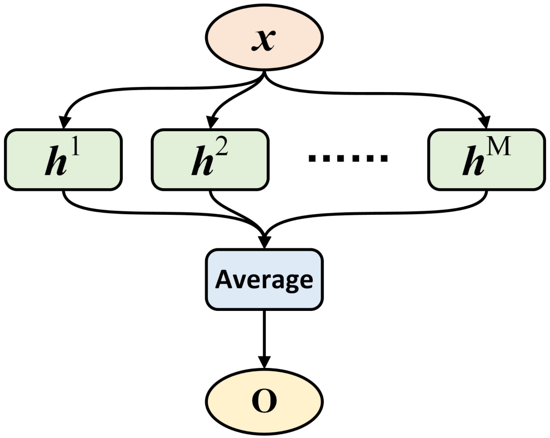


U-Net Architecture [3]

[1] L. Mosser, O. Dubrule, and M. J. Blunt, "Stochastic reconstruction of an oolitic limestone by generative adversarial networks," *CoRR*, vol. abs/1712.02854, 2017. [Online]. Available:http://arxiv.org/abs/1712.02854
[2] "Explain about auto encoder? details about encoder, decoder and bottleneck?," *i2tutorials*, 18-Oct-2019. [Online]. Available: https://www.i2tutorials.com/explain-about-auto-encoder-details-about-encoder-decoder-and-bottleneck/. [Accessed: 14-Oct-2021].
[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

Background: Ensemble Deep Learning

- Combine predictions from multiple NNs (base learners) to (hopefully) make a better final decision
- How to combine (better than simple average)?
  - Weighted average → Stacked Generalization (SG)
- How to determine weights?
  - Just let another ML model (meta-learner) learn it!
- Of course, it's not omnipotent
  - More storage memory, longer inference time



Ensemble Deep Learning [1]

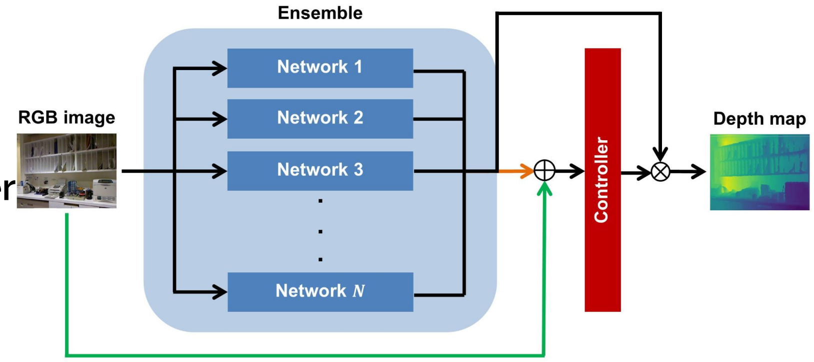
**"It has been applied in various tasks,
but still no application in monocular depth estimation"**

[1] "Introduction," *Ensemble-PyTorch*. [Online]. Available: https://ensemble-pytorch.readthedocs.io/en/latest/introduction.html. [Accessed: 13-Oct-2021].

Research: Objective

"Supervised Monocular Depth Estimation via Stacked Generalization"

- Study SG in monocular depth estimation
- Compare performance of different SG frameworks with simple average (baseline)
  - Should we train base learners and meta-learner **separately** or **simultaneously**?
  - Should we **freeze** or **fine-tune** base-learners when train meta-learner?
  - What should be **inputs of the meta-learner**?
  - How the **performance of base learners** affects the performance of ensemble?



Overview of training pipeline

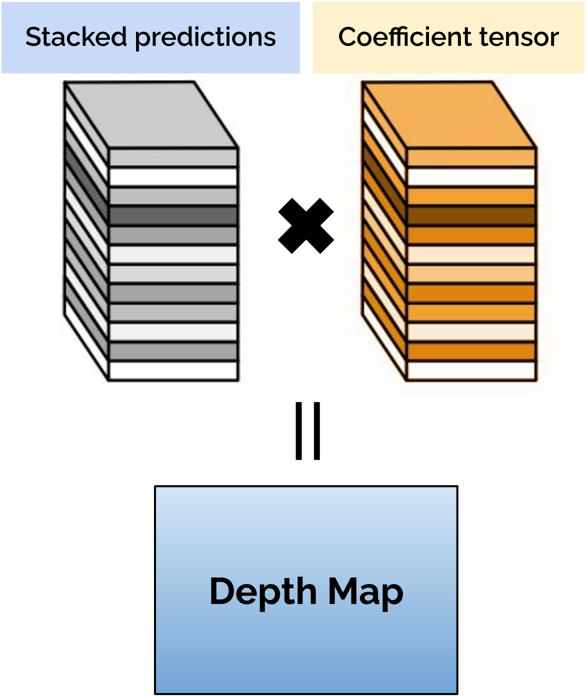
Research: Methodology

Base learners

- Adopted 3 SOTA architectures [1, 2, 3]
- Some modifications to cope with SG's drawbacks and hardware's limitation (1 = lower #param, 2 = lower latency)
  - Employ pretrained GhostNet as encoder (1,2)
  - Use depthwise separable convolutions (1)
  - Interpolating after convolution instead of before (2)

Meta-learner

- U-Net architecture with above modifications
- Inputs are either predictions from base learners or RGB image
- Output are coefficients tensor $[\mathbf{W}]_{ijk}$



[1] S.F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," *CoRR*, vol. abs/2011.14141, 2020. [Online]. Available: https://arxiv.org/abs/2011.14141
[2] J. H. Lee, M. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," *CoRR*, vol. abs/1907.10326, 2019. [Online]. Available: http://arxiv.org/abs/1907.10326
[3] M. Song, S. Lim, and W. Kim, "Monocular depth estimation using laplacian pyramid-based depthresiduals," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2021.

Research: Methodology

Loss Functions

- (1) Pixel-wise depth loss [1]
  - Mitigate pixel-wise difference
  - Human perceive logarithmically
  - Uses when train every model
- (2) Bin center density
  - Bichamfer Loss [2]
  - Encourage distribution of bin centers to follow distribution of ground truth depth values
  - Uses in Adabins only

$$L_{\text{total}} = L_{\text{pixel}} + L_{\text{bins}}$$

$$L_{\text{pixel}} = \alpha \sqrt{\frac{1}{N}\sum_{i=1}^{N} y_i^2 - \frac{\lambda}{N^2}\Big(\sum_{i=1}^{N} y_i\Big)^2}$$

where $y_i^2 = \log(d_i) - \log(d_i^*)$ and d_i^* is ground truth depth

$$L_{\text{bins}} = \textbf{BiChamfer}(c(b), D) + \textbf{BiChamfer}(D, c(b))$$

Note that $\lambda = 0.85$ and $\alpha = 10$ are used as same as the original Adabins

Bichamfer Loss

$$d_{CD}(S_1, S_2) = \sum_{x \in S_1} \min_{y \in S_2} \|x - y\|_2^2 + \sum_{y \in S_2} \min_{x \in S_1} \|x - y\|_2^2$$

[1] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *CoRR*, vol. abs/1406.2283, 2014. [Online]. Available: http://arxiv.org/abs/1406.2283
[2] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3d object reconstruction from a single image," *CoRR*, vol. abs/1612.00603, 2016. [Online]. Available: http://arxiv.org/abs/1612.00603

Research: Experiment

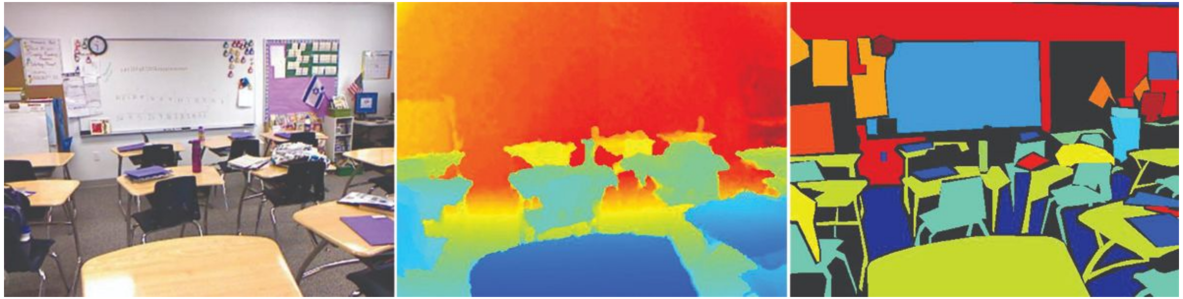
Implementations

- Implemented in Pytorch, trained in a Laboratory's server using distributed training
  - Intel(R)Xeon(R) CPU E5-2534 @ 3.40GHz with 256 GB of RAM
  - 10 NVIDIA GeForce GTX 1080 Ti GPUs with 12 GB memory
- Train with AdamW optimizer following 1-cycle policy for fast convergence
  - Maximum learning rate (lr) for each model is determined from lr range test
  - Linear warm-up for 30% of iteration from lr/25, followed by cosine annealing to lr/100
- Batch size 32, Weight decay 1e-4
- Other hyperparameters are tuned via grid search and random search
- Monitored using Weights and Biases platform
- NOT employing bootstrap since lower #data might affect performance of model
  - One base learner uses a Visual Transformer (ViT) which is extremely data hungry

Research: Experiment

Dataset

- 464 different indoor scenes
  - Official split → 249 training and 215 for testing (654 images)
- Monocular video sequences of scenes & ground truth depth from RGB-D camera
- Operation frequency of RGB and Depth camera are different
  - 120K image-depth pairs are sampled and matched → 24,231 training samples [1]



A sample of raw image, preprocessed depth, and labeled from the dataset [1]

[1] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from rgbd images,” in *Computer Vision – ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, andC. Schmid, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 746–760.

Research: Experiment

Data augmentation

- Data augmentation refers to techniques to prevent overfitting by generating more (feasible) training examples from original data
- Follow data augmentation techniques described in [1]:
    - Random horizontal flipping with probability of 0.5
    - Random contrast, brightness, and color adjustment in a range of [0.9, 1.1] with probability of 0.5
    - Random crop of size 416 ✖ 544
    - Random rotation of degree in a range of [-2.5,2.5]

[1] S.F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," *CoRR*, vol. abs/2011.14141, 2020. [Online]. Available: https://arxiv.org/abs/2011.14141

Research: Experiment

Evaluation metrics

- **Threshold Accuracy**: % of d_i s.t. $\max\left(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}\right) = \delta <$ threshold, usually threshold $= 1.25, 1.25^2, 1.25^3$
- **Average Relative Error (REL)**: $\frac{1}{N}\sum\left(\frac{|d_i - d_i^*|}{d_i^*}\right)$
- **Root Mean Squared Error (RSME)**: $\sqrt{\frac{1}{N}\sum(d_i - d_i^*)^2}$
- **Average** $\log_{10}$ **Error**: $\frac{1}{N}\sum|\log_{10}(d_i) - \log_{10}(d_i^*)|$
- **Squared REL (Sq REL)**: $\frac{1}{N}\sum\frac{\|d_i - d_i^*\|}{d_i^*}$
- **RSME of logarithm (RSME log)**: $\sqrt{\frac{1}{N}\sum\|\log d_i - \log d_i^*\|^2}$

[1] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *CoRR*, vol. abs/1406.2283, 2014. [Online]. Available: http://arxiv.org/abs/1406.2283

Research: Result

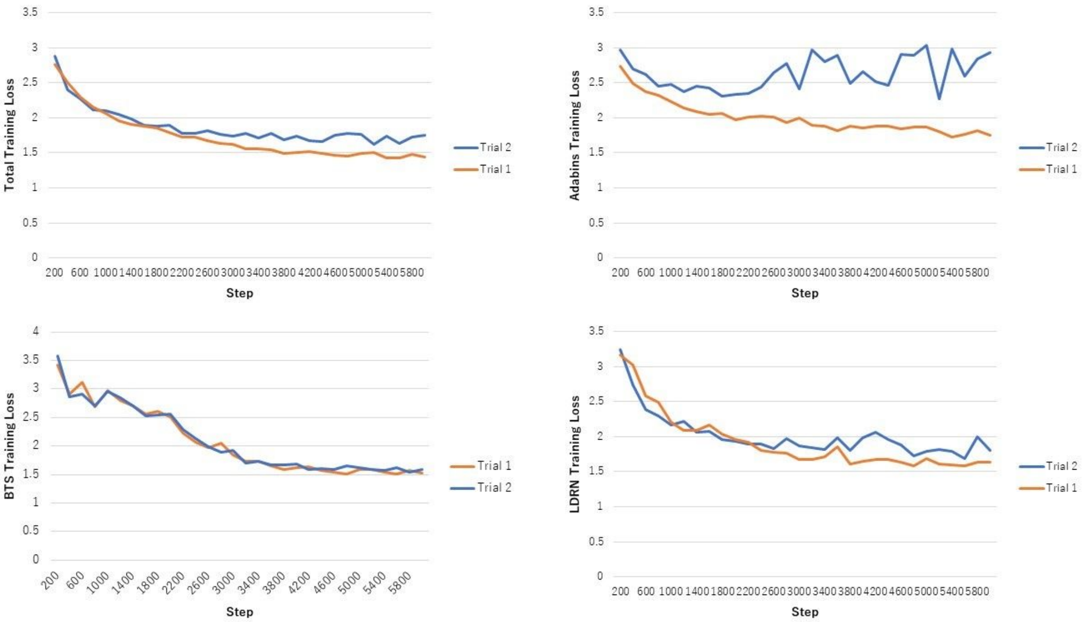
- Using RGB image in meta-learner → worse REL, but better overall performance
  - Likely caused by insufficient representation capability of meta-learner

Variant	#Params	*higher is better*			*lower is better*				
		$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	REL	Sq REL	RSME	RSME log	log10
Base: Adabins	17.2M	0.8106	0.9641	0.9919	0.1463	0.0875	0.5019	0.1788	0.0604
Base: LDRN	14.9M	0.8306	0.9661	0.9925	0.1320	0.0875	0.4561	0.1675	0.0564
Base: BTS	8.9M	0.8567	0.9724	0.9932	0.1202	0.0749	0.4326	0.1558	0.0521
Baseline	-	0.8564	**0.9758**	**0.9948**	0.1216	0.0739	**0.4261**	0.1537	0.0518
SG: Simultaneous	4.3M	0.8538	0.9727	0.9935	0.1199	0.0741	0.4340	0.1553	0.0521
SG: RGB, Tuned	4.3M	0.8581	0.9746	<u>0.9944</u>	0.1210	0.0733	0.4274	0.1540	0.0517
SG: RGB, Freeze	4.3M	0.8578	<u>0.9748</u>	<u>0.9944</u>	0.1195	<u>0.0727</u>	0.4290	0.1538	<u>0.0516</u>
SG: D	4.3M	<u>0.8590</u>	0.9745	0.9941	**0.1189**	0.0728	<u>0.4267</u>	<u>0.1535</u>	**0.0514**
SG: RGB + D	4.3M	**0.8595**	<u>0.9748</u>	<u>0.9944</u>	<u>0.1193</u>	**0.0724**	<u>0.4267</u>	**0.1533**	**0.0514**

Table 1: Evaluation results on NYU Depth V2. Bold and underline denote the first and second place, respectively. The proposed method (*SG: RGB + D*) yields competitive results on all metrics.

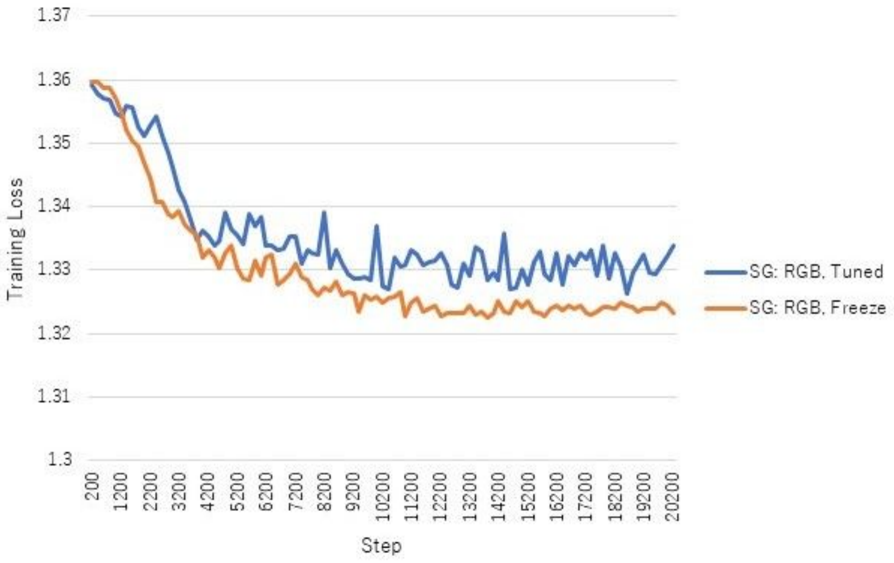
Research: Result (2)

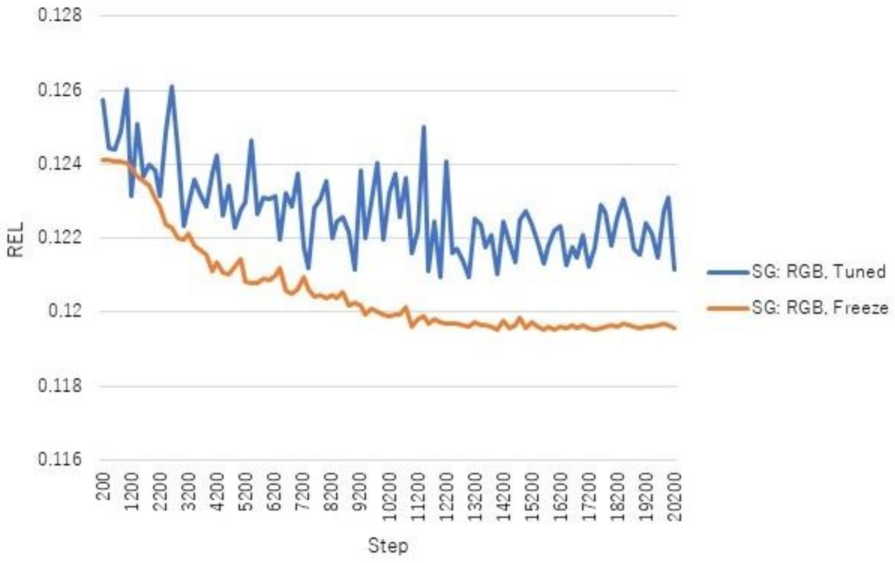
- Training base learners and meta-learner together is **unstable**
  - Likely caused by performance gap among base learners
  - Require careful hyperparameter tuning to ensure convergence



Research: Result (3)

- Fine-tuning base learners when train meta-learner yields <u>**useless**</u> loss fluctuation





Research: Result (4)

- **Computational resource required**
  - No significant difference among meta-learner variants

Variants	Param Size (MB)	Total Mul-adds (G)	Training time (hrs)	Inference time (fps)
Base: Adabins	68.90	9.57	~6	15.94
Base: BTS	35.60	12.07	~8	9.57
Base: LDRN	59.57	14.83	~5	18.76
Baseline	-	-	-	6.5
SG: D	17.06	1.66	~11	5.86 (ensemble)
SG: RGB + D	17.07	1.69	~11	5.9 (ensemble)

Research: Future works

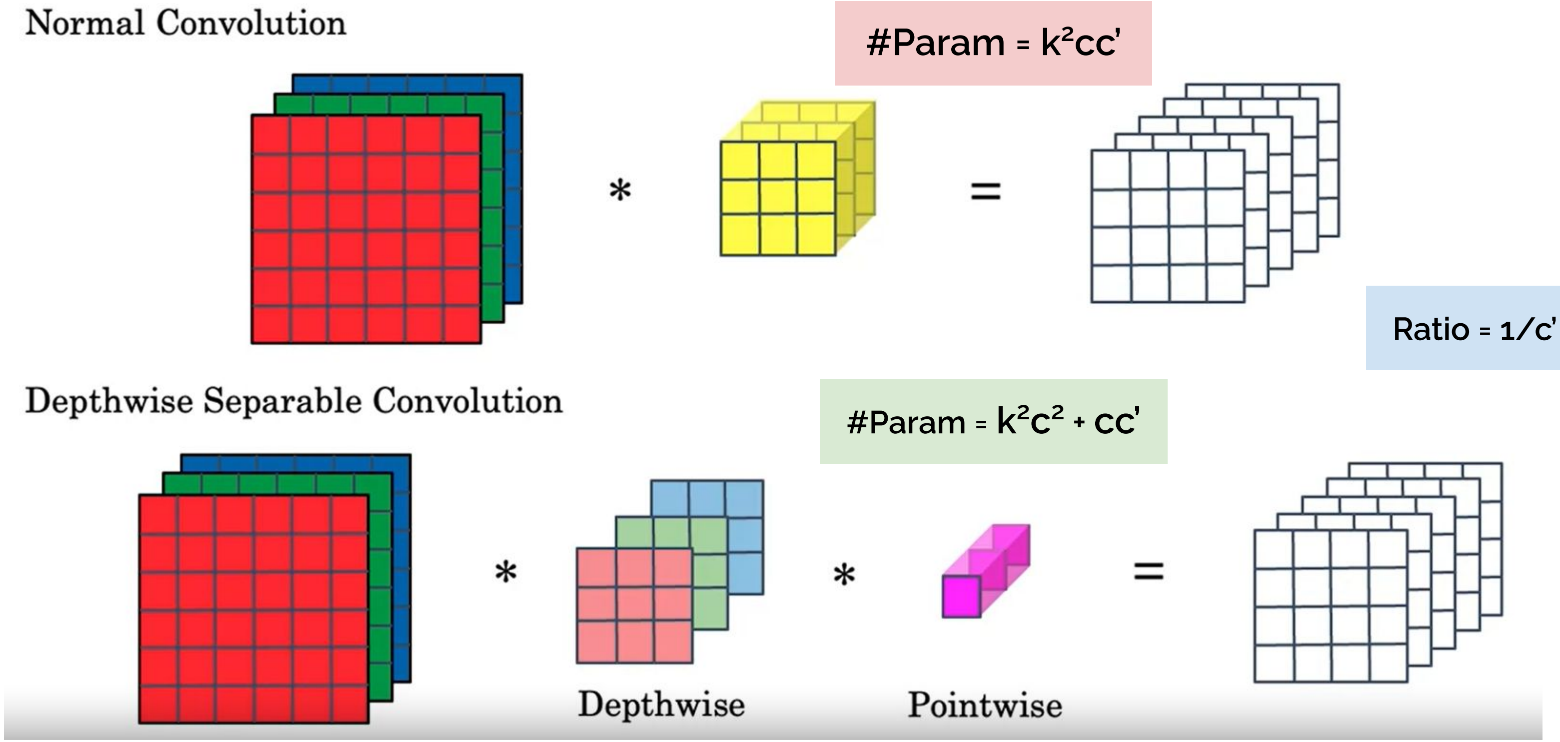
Current Plan

- Experiment with
  - Base learners without significant performance gap
  - Different/larger encoders with more representation capability
- Qualitative results

Need Advices!

- **Should I repeat the same experiment several times and average their results?**
- **Should I try bootstrap? How to compare the result with those w/o bootstrap?**
- **Any advices is welcome!**

Appendix: Depthwise Separable Convolution



#Param = k^2cc'

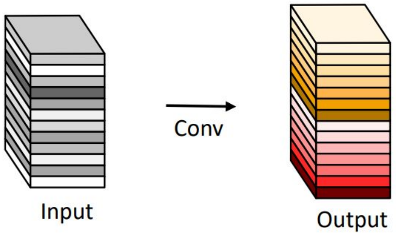
Ratio = $1/c' + 1/k^2$

#Param = $k^2c^2 + cc'$

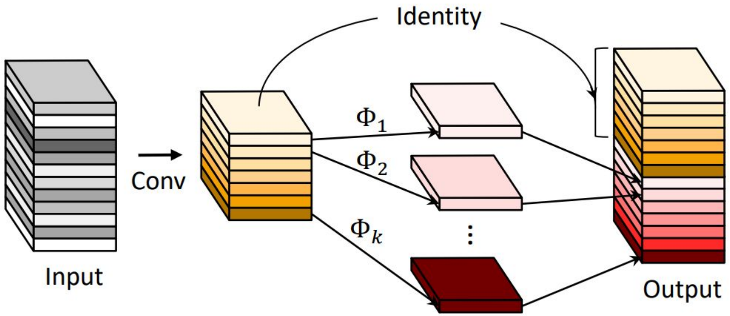
[1] “Convolutional Neural Networks,” Coursera. [Online]. Available: https://www.coursera.org/learn/convolutional-neural-networks/home/welcome. [Accessed: 15-Oct-2021].

Appendix: GhostNet

- Based on observation that the output feature maps of convolutional layers often contain much <u>**redundancy**</u>
- Generate some feature maps through usual convolution. Then, apply linear operations to generate more feature maps
- GhostNet is ghost modules arranged in a structure similar to MobileNetV2



(a) The convolutional layer.

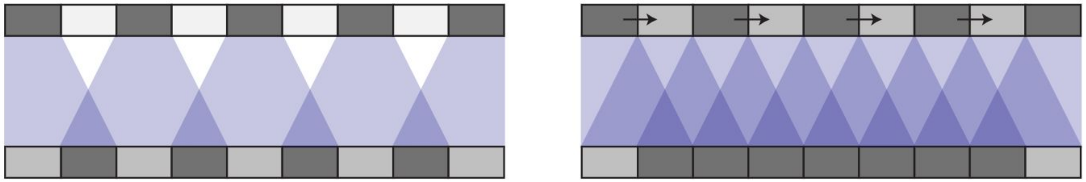


(b) The Ghost module.

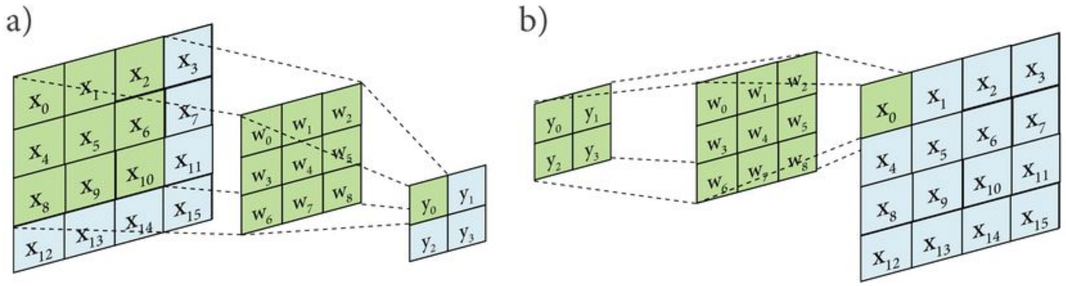
[1] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheapoperations," *CoRR*, vol. abs/1911.11907, 2019. [Online]. Available: http://arxiv.org/abs/1911.11907

Appendix: Why not use transposed convolution?

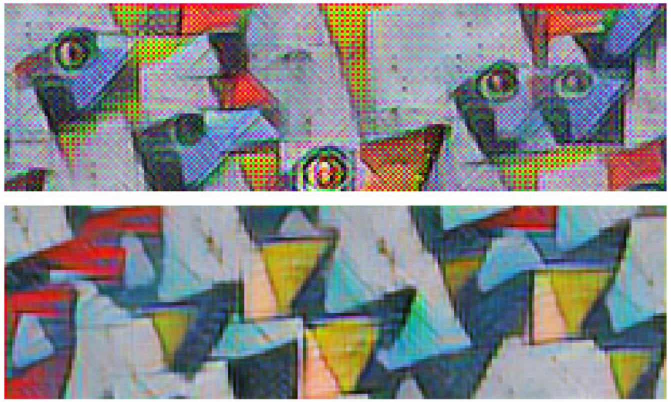
- Transposed convolution generates **checkerboard pattern**



Coverage of Transposed Conv and Conv [1]



Conv and Transposed Conv [2]

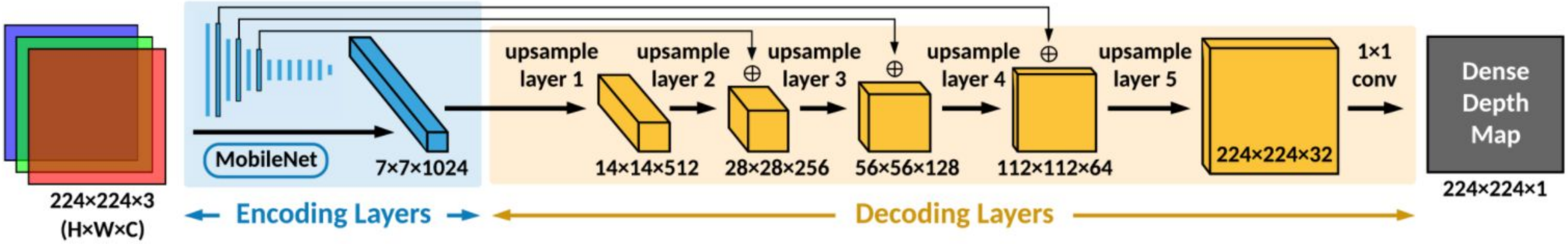


Checkerboard pattern [1]

[1] A. Odena, V. Dumoulin, and C. Olah, “Deconvolution and checkerboard artifacts,” *Distill*, 17-Oct-2016. [Online]. Available: https://distill.pub/2016/deconv-checkerboard/. [Accessed: 15-Oct-2021].
[2] L. Mosser, O. Dubrule, and M. J. Blunt, “Stochastic reconstruction of an oolitic limestone by generative adversarial networks,” *CoRR*, vol. abs/1712.02854, 2017. [Online]. Available:http://arxiv.org/abs/1712.02854

Appendix: FastDepth

- This paper proposes a lightweight decoder for monocular depth estimation
    - Contains only **depthwise separable convolutions**
    - When decoding, interpolate after convolution instead of before
- While it has low-latency and smaller size (even smaller after **network pruning**), its accuracy is naturally worse than SOTAs



[1] D. Wofk, F. Ma, T. Yang, S. Karaman, and V. Sze, "Fastdepth: Fast monocular depth estimation on embedded systems," *CoRR*, vol. abs/1903.03273, 2019. [Online]. Available:http://arxiv.org/abs/1903.03273