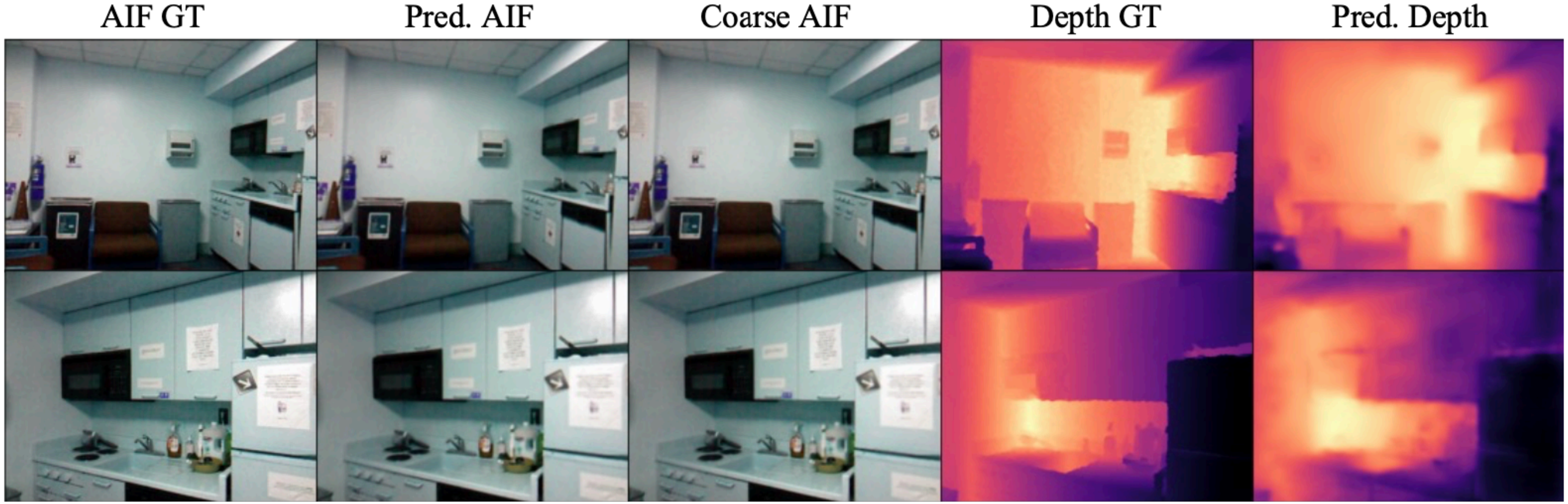
上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Fully Self-Supervised Depth Estimation from Defocus Clue

CVPR 2023

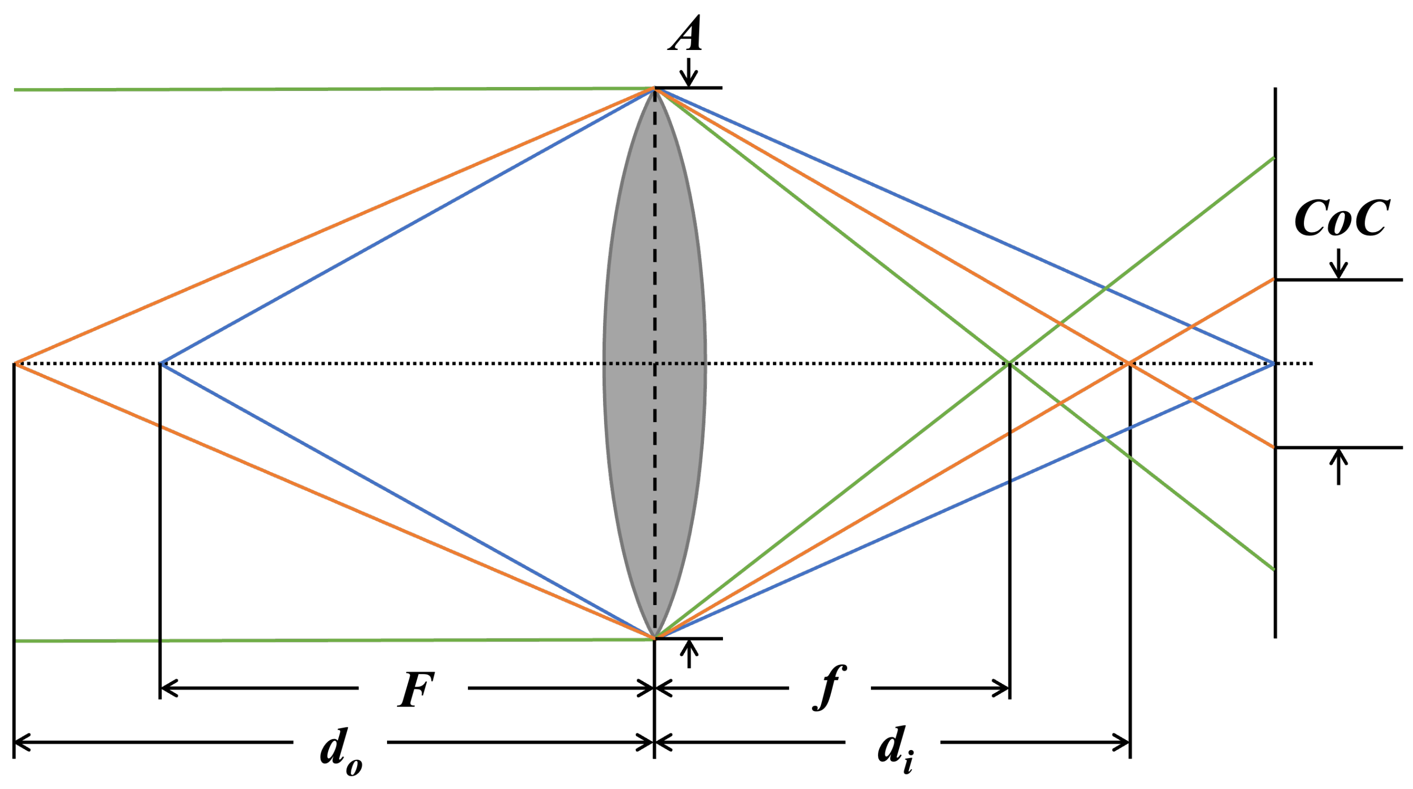


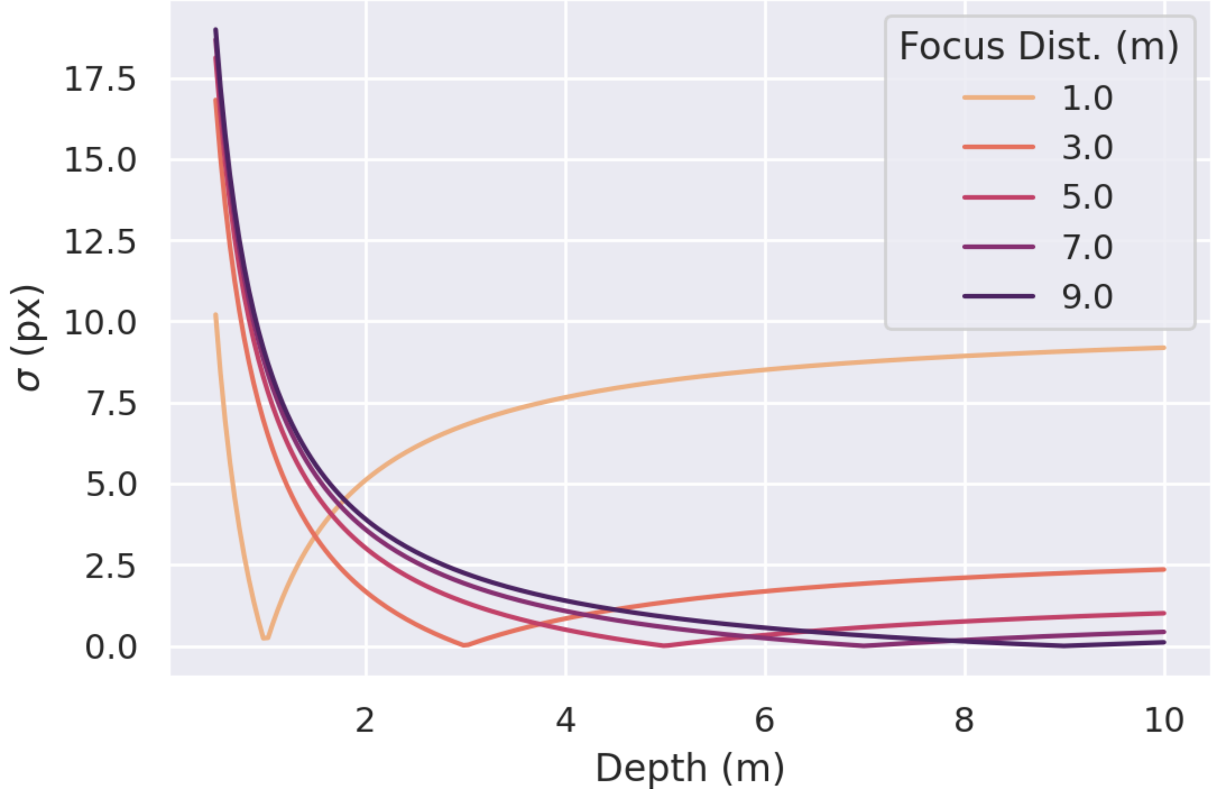
Haozhe Si*, Bin Zhao*, Dong Wang†, Yunpeng Gao, Mulin Chen, Zhigang Wang, Xuelong Li†

Introduction: Depth-from-Defocus (DFD)

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- DFD estimates the depth base on the degree of blur or defocus in the image.
  - Objects at different distances from the camera will be blurred or defocused to different extents.
- Problems of existing DFD works:
  - <u>Supervised DFD</u> works require accurate depth ground-truth;
  - <u>Self-supervised DFD</u> methods rely on all-in-focus (AIF) during training. However, capturing AIF images can be challenging in real-world scenarios.

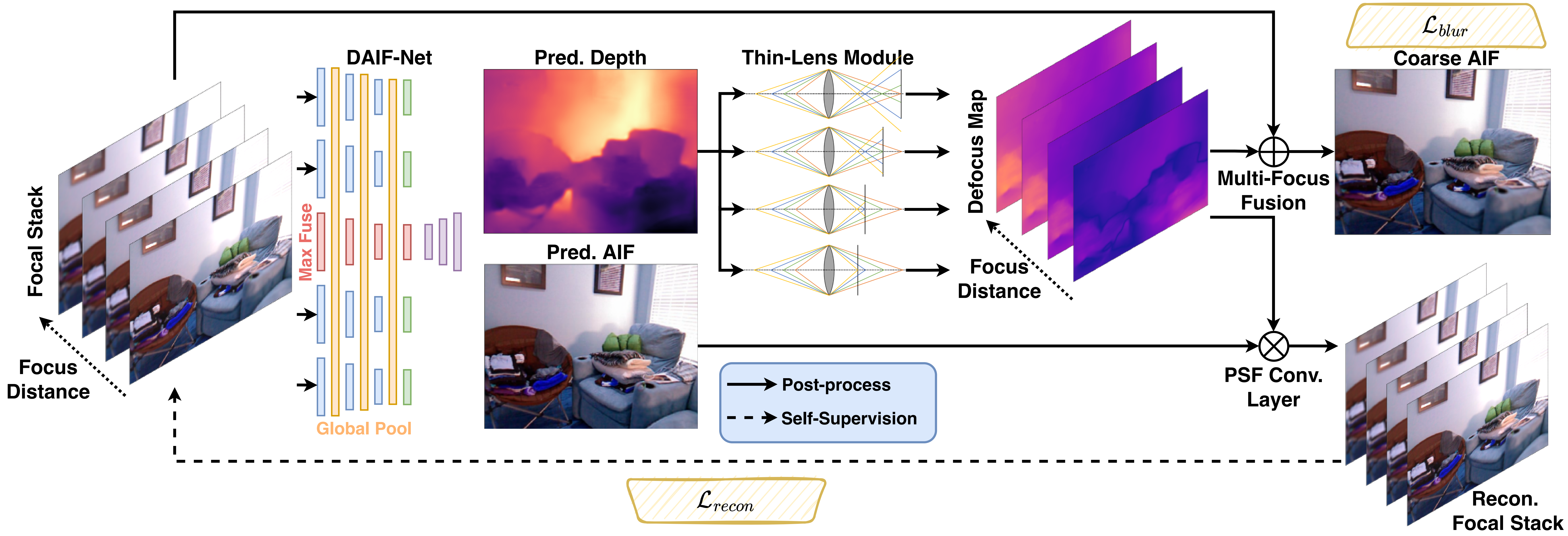




Problem Statement: Fully Self-supervised DFD

- **A more realistic setting:**
  - The availability of depth and AIF images ground-truth is deprived.
  - Only focal stacks are provided in model training.
  - Fully Self-supervised DFD!
- **Challenge:**
  - No longer have the direct/indirect optimization goal for the model.
- **Benefits:**
  - More practical data collection process in real-world scenario;
  - More flexible and adaptable training, better generalization taking advantage from SSL.

Proposed Framework



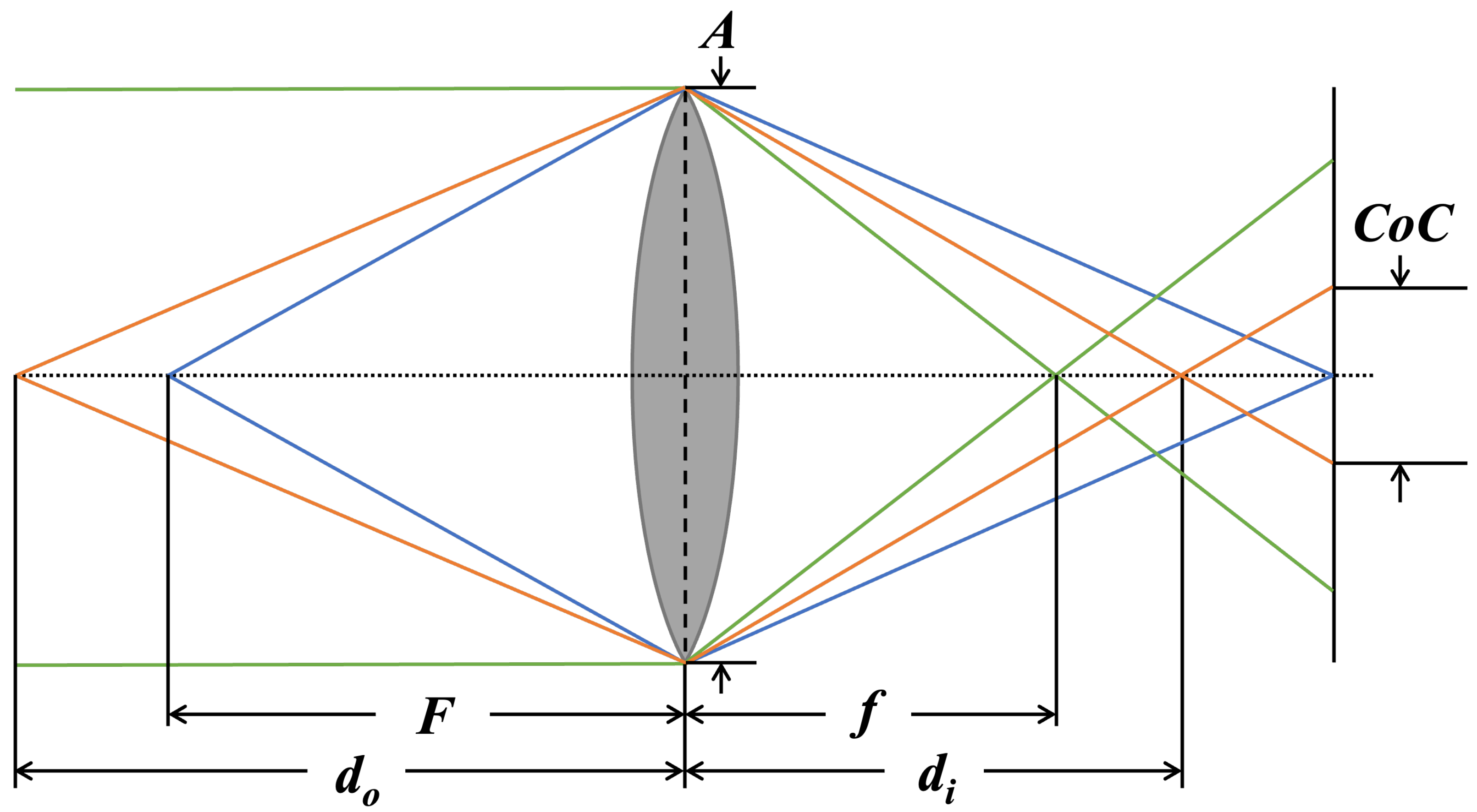
- **Input:** Sparse focal stack.
- **DAIF-Net:** Predicts depth and AIF images from input.
- **Optical Model for Defocus Blur:** Render defocus image from predicted depth and AIF images.
- **Self-supervision:** Reconstruct the input focal stack.

Optical Model for Defocus Blur

- **Defocus Map Generation:**
  - Use defocus map to quantitatively measure the defocus blur in an image.
  - The circle of confusion (*CoC*) measures the diameter of such a blurry circle.
  - The relationship between *CoC* and depth is well established by the thin-lens model.
  - We further measure the radius of the *CoC* in pixels:

$$\sigma = \frac{CoC}{2 \cdot p} = \frac{1}{2p} \frac{|d_o - F|}{d_o} \frac{f^2}{N(F - f)}$$

  - We can generate defocus map from the predicted depth map with above equation.



Optical Model for Defocus Blur

- **Defocus Image Rendering:**
  - Using Point Spread Function (PSF)
    - In practice, we use a simplified disc-shaped PSF, *i.e.*, a Gaussian kernel:

$$\mathcal{F}_{x,y}(u,v) = \frac{1}{2\pi\Sigma_{x,y}^2}\exp\left(-\frac{u^2+v^2}{2\Sigma_{x,y}^2}\right)$$

    - The defocus image is produced by convolving AIF images with PSF:

$$J := I \otimes \mathcal{F}$$

- **Why Optical Model?**
  - <u>Physically explainable process:</u> Encourages the predicted depth and AIF images to have their corresponding physical properties.
  - <u>Optimization issue:</u> Using an optical model requires no training.
  - <u>Robust and provable performance</u>

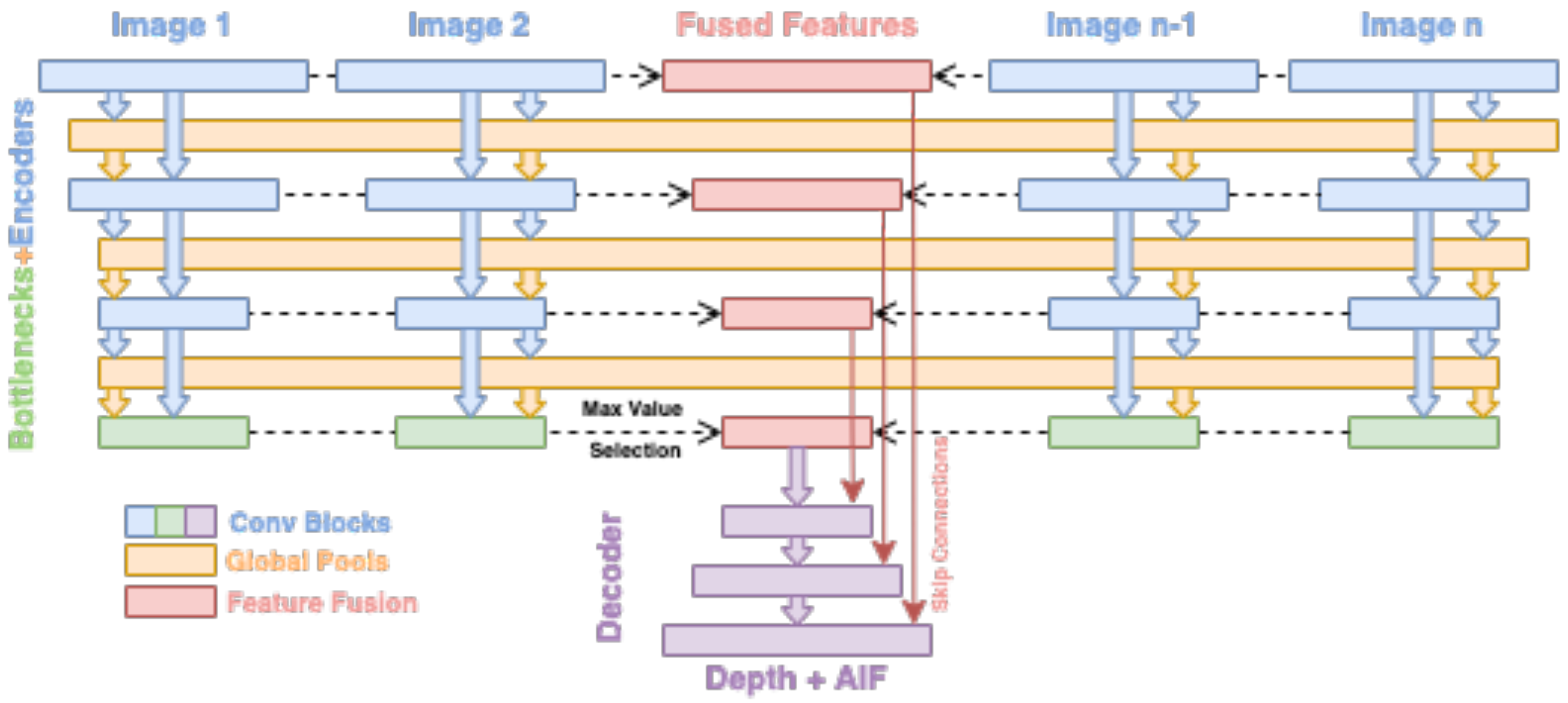
DAIF-Net

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- **Motivation:**
  - The optical model depicts the forward process of generating defocus image from the AIF image and the depth:
  $$J_F = \mathscr{G}_F(I, D).$$
  - The DAIF-Net predicts depth and AIF image from the focal stack, which is the reverse process of previous equation:
  $$\{I, D\} = \mathscr{D}(J_F)$$
  - **Ill-posed problem:** calculating **two** variables from **one** input.
  - **Solution:** Taking advantage from the focal stack:
  $$\{I, D\} = \mathscr{D}(J_F^0, J_F^1, \cdots, J_F^k)$$
    - The accurate depth map and AIF image is the only solution to this equation.

DAIF-Net

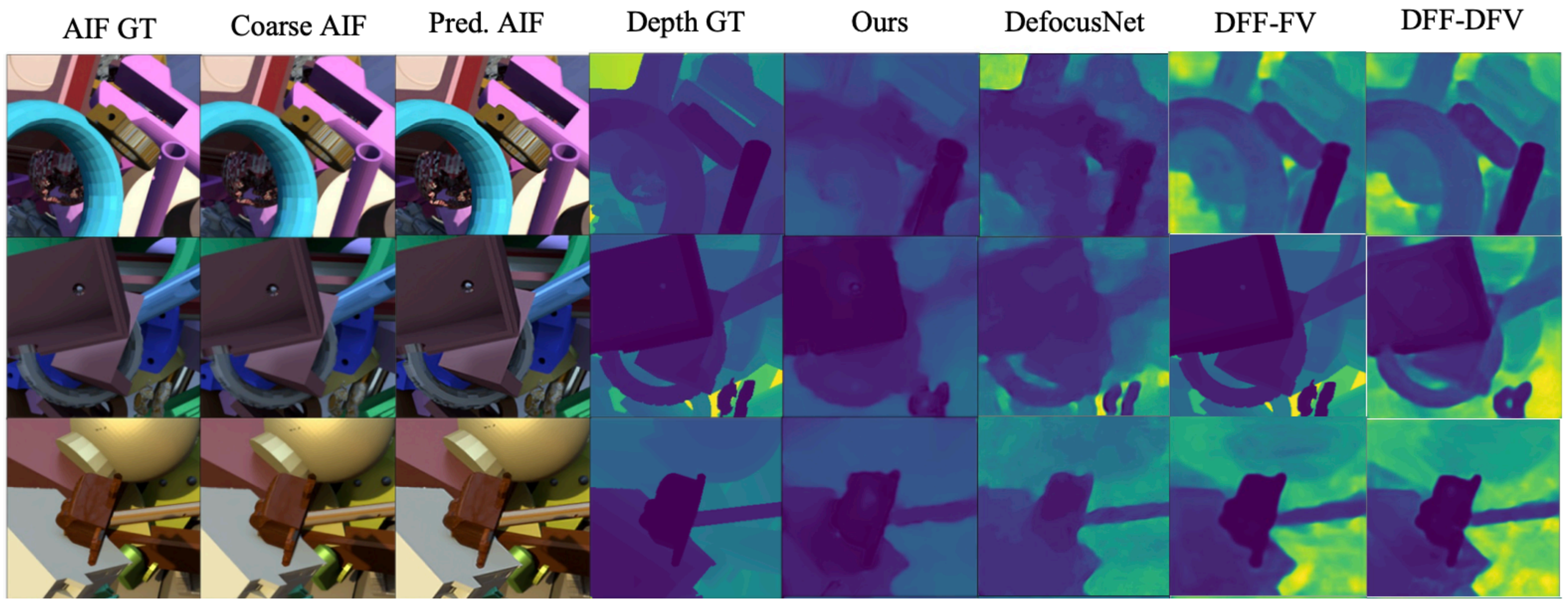
- **Goal:** Estimate the depth map and the AIF image from a focal stack.
- **Architecture:**
  - Modified U-Net encoder:
    - Each image pass through the same encoder and bottleneck.
    - The features will be fused across the inputs for decoding and skip connections.
  - Layer-wise global pooling:
    - Leverage sharpness as the link between depth and AIF.
    - Emphasize sharper regions in the multi-input encoder.



Empirical Results: Synthetic Dataset

Re-rendered DefocusNet Dataset

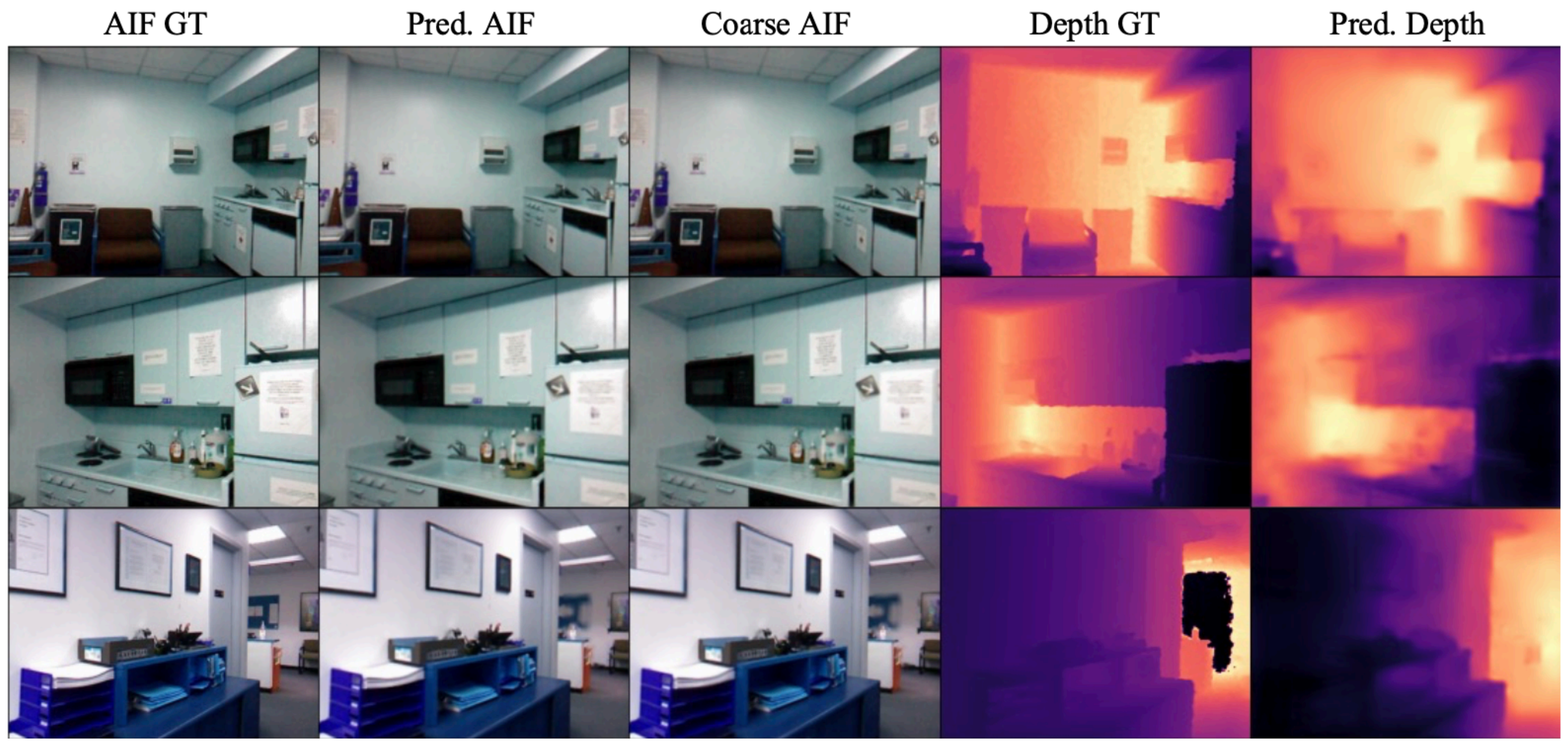
Methods	Regular					0.5m				
	$\delta1\uparrow$	$\delta2\uparrow$	$\delta3\uparrow$	RMSE$\downarrow$	AbsRel$\downarrow$	$\delta1\uparrow$	$\delta2\uparrow$	$\delta3\uparrow$	RMSE$\downarrow$	AbsRel$\downarrow$
Supervised Learning										
DefocusNet [15]	0.912	0.967	0.983	0.194	0.090	0.911	0.933	0.938	0.062	0.069
DFF-FV [27]	0.883	0.953	0.980	0.231	0.107	0.977	0.996	0.999	0.023	0.032
DFF-DFV [27]	0.921	0.977	0.990	0.219	0.104	0.976	0.996	0.999	0.023	0.031
Self-supervised Learning										
Ours	0.746	0.883	0.938	0.351	0.177	0.889	0.987	0.992	0.072	0.138



Empirical Results: Synthetic Defocus Blur

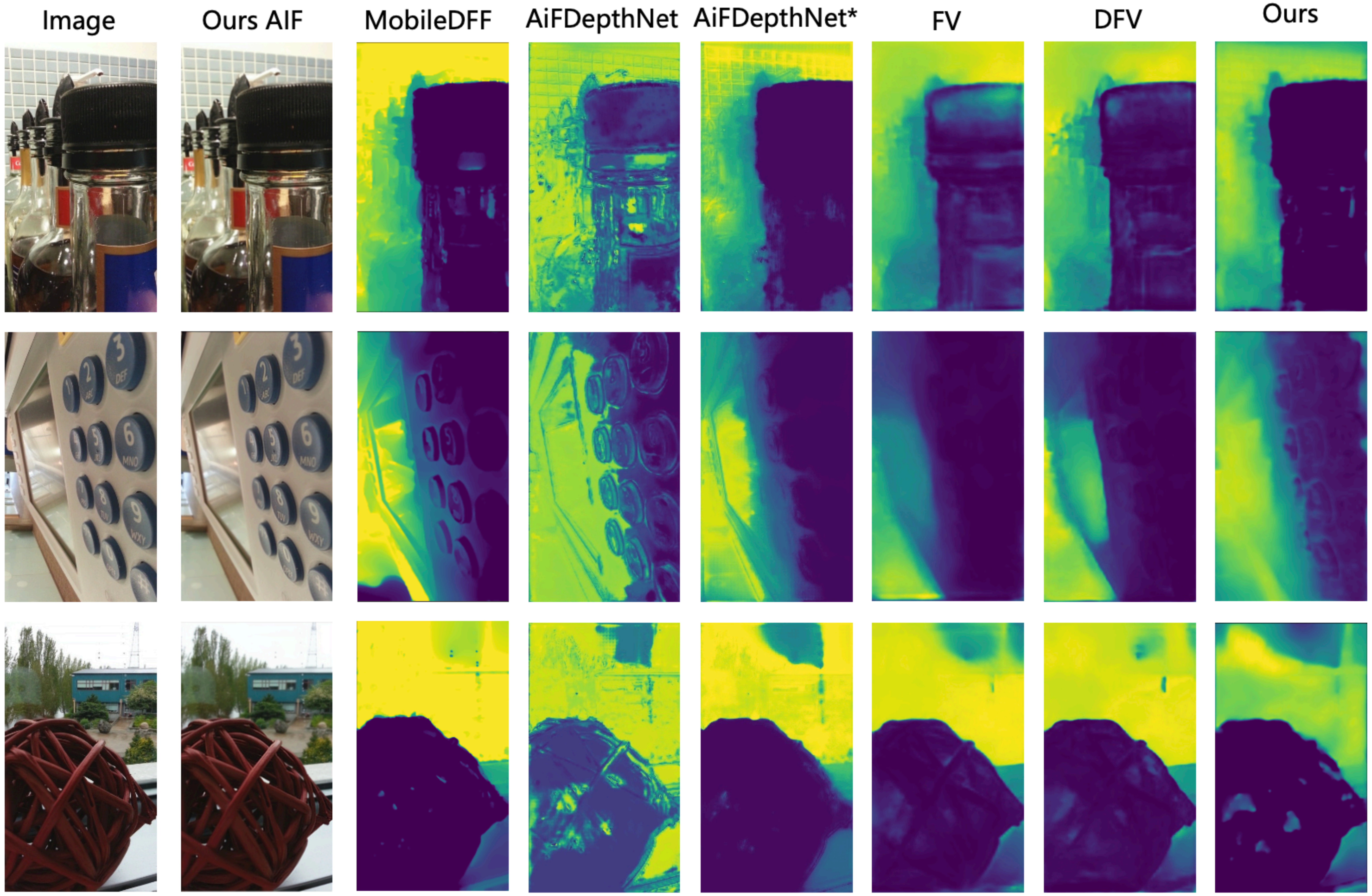
NYUv2

Methods	Input	$\delta1\uparrow$	$\delta2\uparrow$	$\delta3\uparrow$	RMSE$\downarrow$	AbsRel$\downarrow$
Analytical Methods						
Moeller et al. [16]	focal stack	0.670	0.778	0.912	0.985	0.263
Suwajanakorn, Hernandez, and Seitz [22]	focal stack	0.688	0.802	0.917	0.950	0.250
Self-sup w/ AIF						
Gur and Wolf [8]	in-focus	0.720	0.887	0.951	0.649	0.184
Defocus-Net [14]	defocus	0.732	0.887	0.951	0.623	0.176
Focus-Net [14]	focal stack	0.748	0.892	0.949	0.611	0.172
Supervised Learning						
DFF-FV [27]	focal stack	0.956	0.979	0.988	0.285	0.470
DFF-DFV [27]	focal stack	0.967	0.980	0.990	0.232	0.445
Fully Self-Supervised Learning						
Ours	focal stack	0.950	0.979	0.987	0.325	0.170



Empirical Results: Real Focal Stack Dataset

Mobile Depth



Conclusion

- **Limitation:**
  - Our framework is more suitable for closed scenes and textured scenes, where the defocus blurs are easier to be observed.
- **Contribution:**
  - We design a more realistic and challenging scenario for the DFD, where only focal stacks are available in model training and evaluation.
  - We propose the first completely self-supervised framework for DFD. The framework predicts depth and AIF images simultaneously from a focal stacks and is supervised by reconstructing the input.
  - Our framework performs favorably against the supervised state-of-the-art methods, providing a strong baseline for future self-supervised DFD tasks.

Thanks for Watching!

Scan for More Information