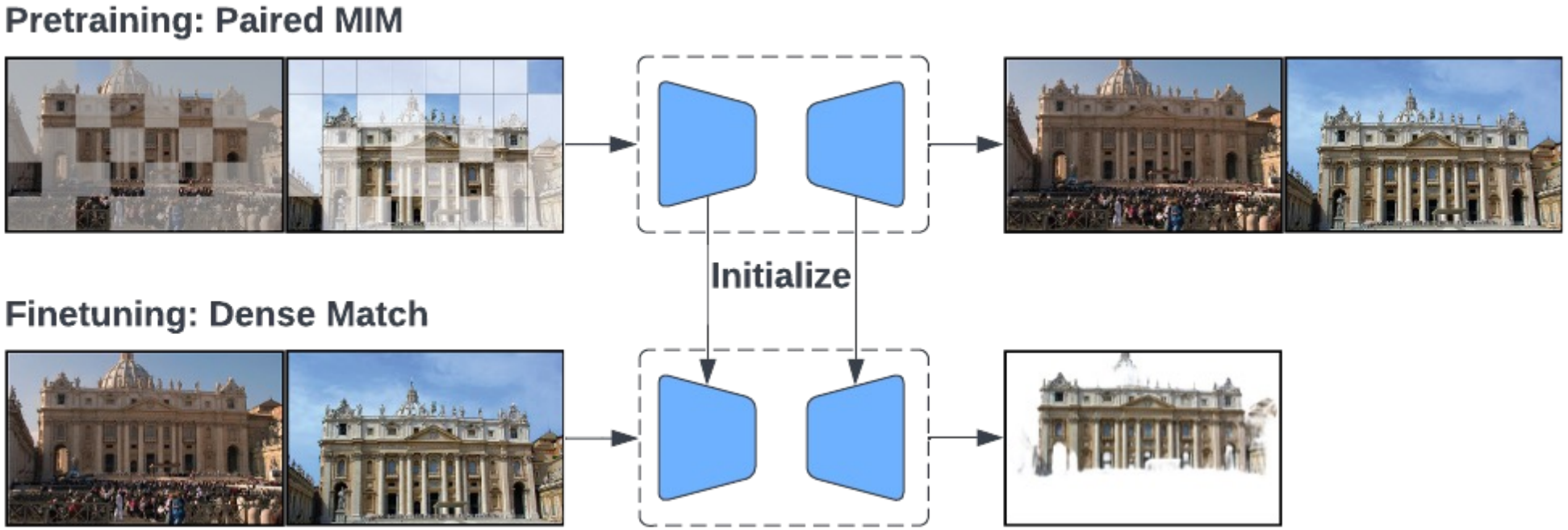
PMatch:
Paired Masked Image Modeling for Dense Geometric Matching

Shengjie Zhu and Xiaoming Liu

Michigan State University

Paper ID:8490
Paper Session: THU-PM-124

PMatch:
Paired Masked Image Modeling for Dense Geometric Matching

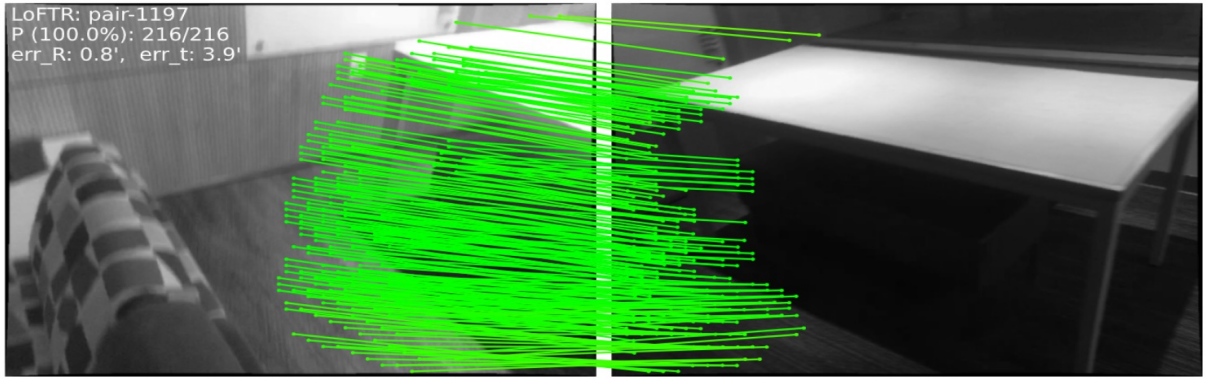


- We propose a Transformer Based Network for Dense Geometric Matching
- We propose paired masked image modeling to pretrain the transformer

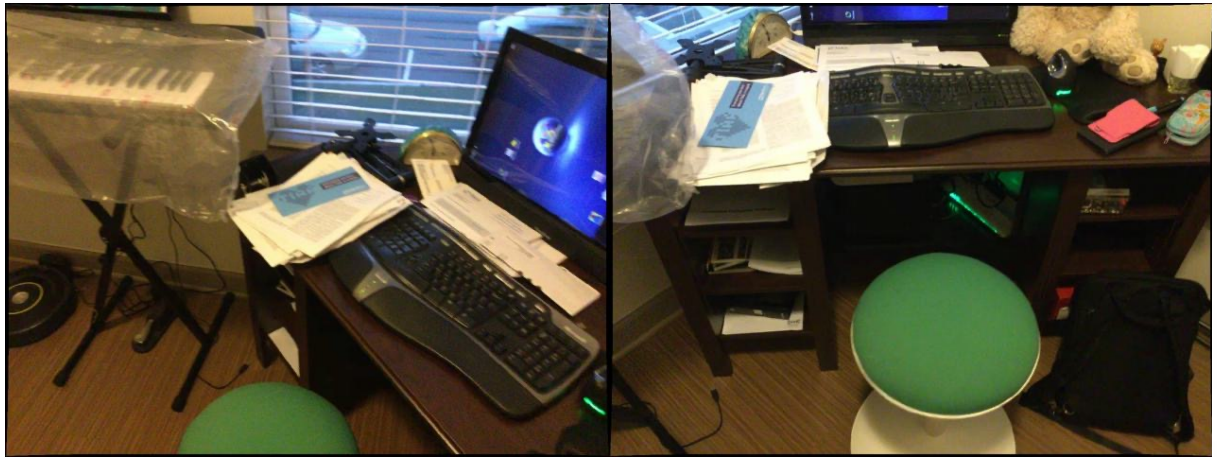
Introduction

Dense and Sparse Correspondence Estimation

- Sparse Correspondence Estimation



- Dense Correspondence Estimation



- Dense Correspondence

- Dense Confidence

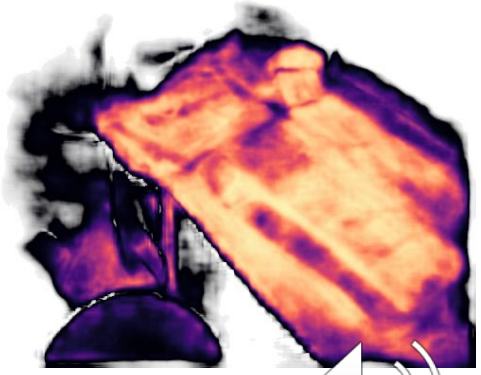
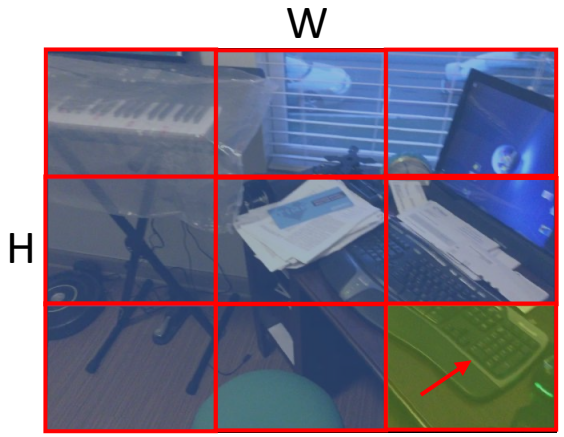
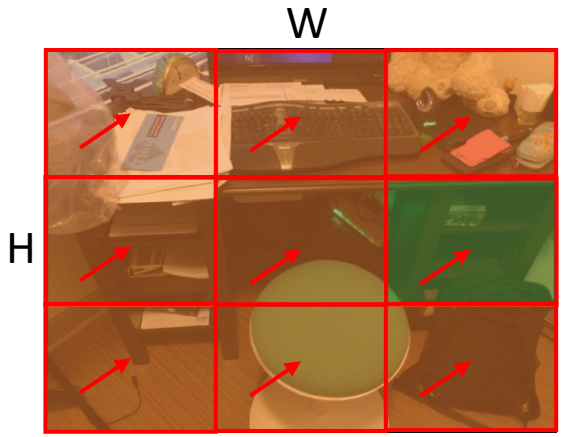
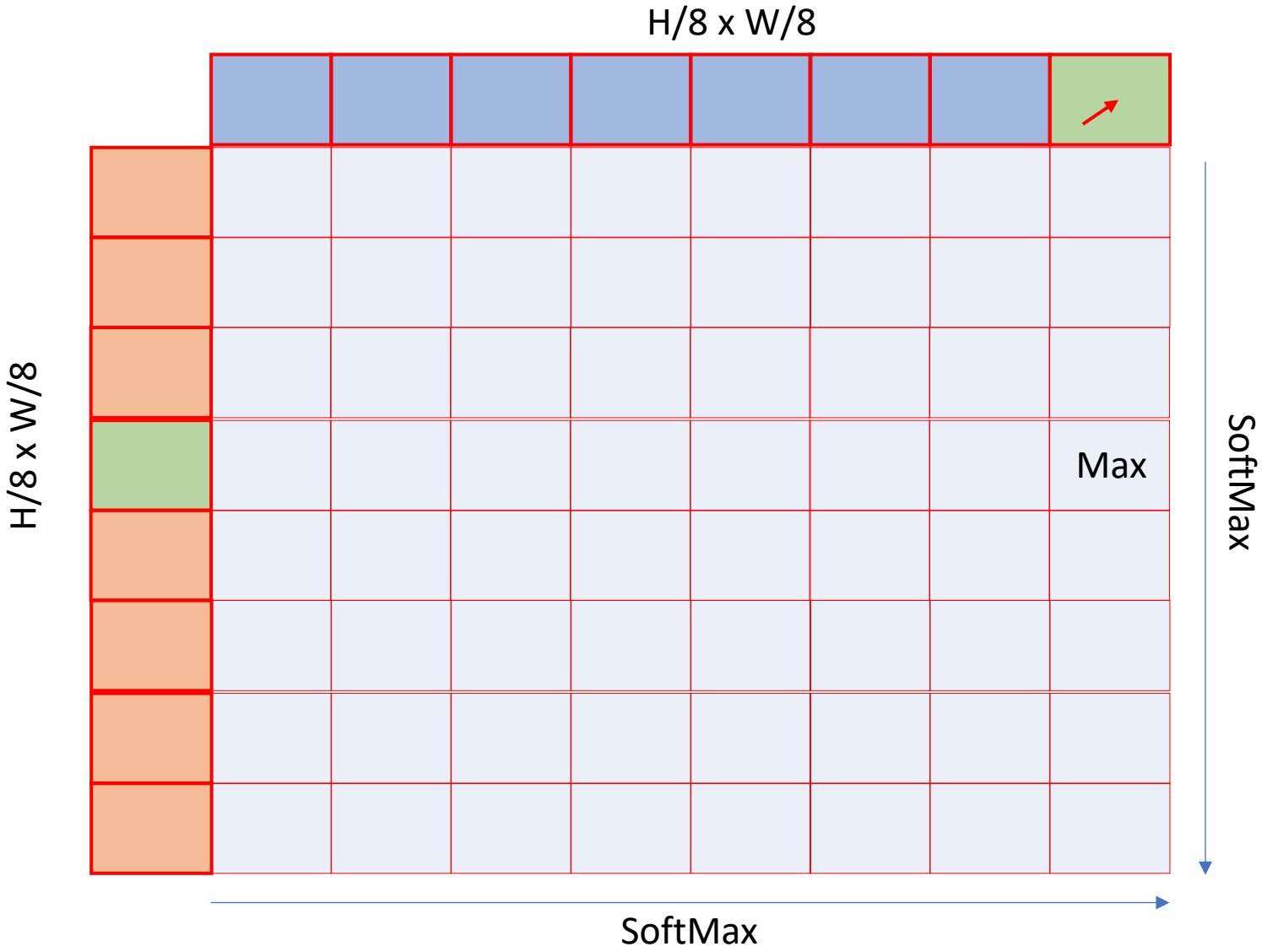


Image Courtesy: https://openaccess.thecvf.com/content/CVPR2021/html/Sun_LoFTR_Detector-Free_Local_Feature_Matching_With_Transformers_CVPR_2021_paper.html

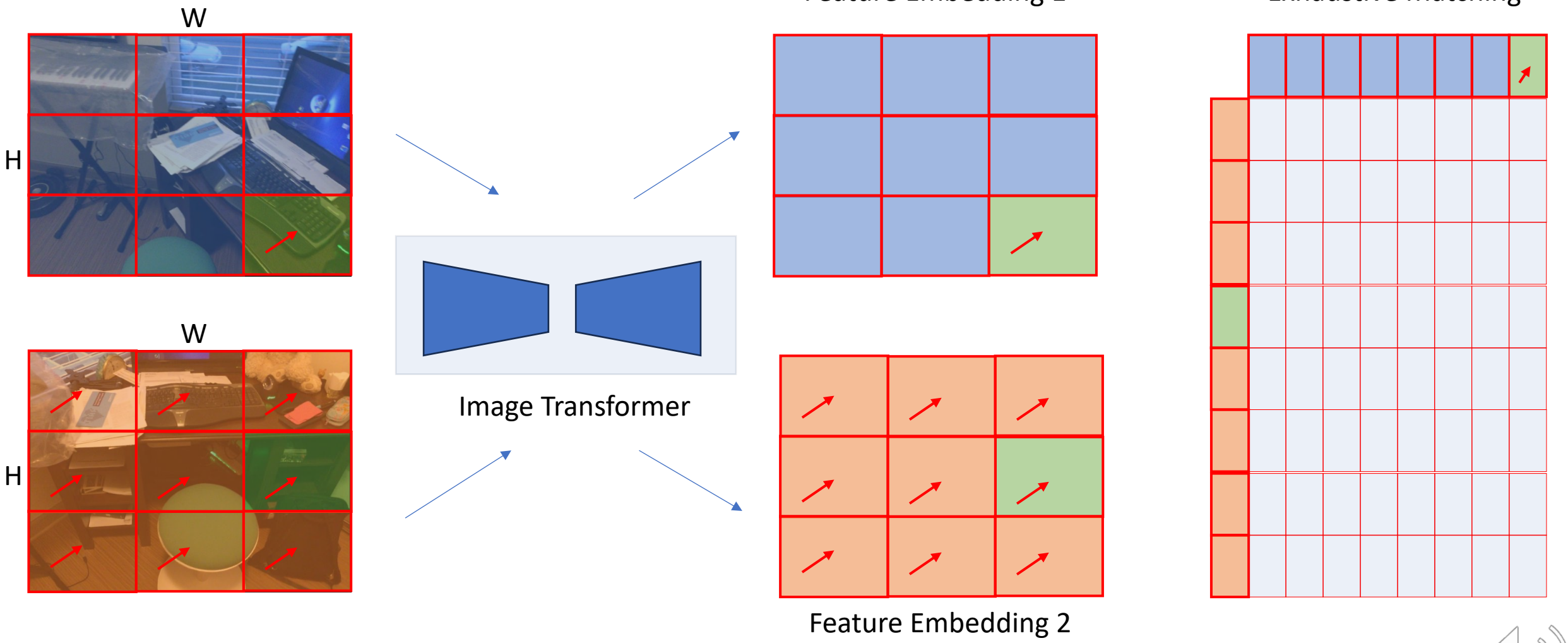
How to Perform Dense Coprrespondence?



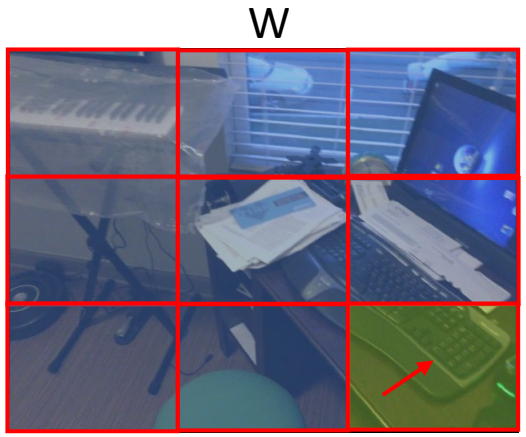


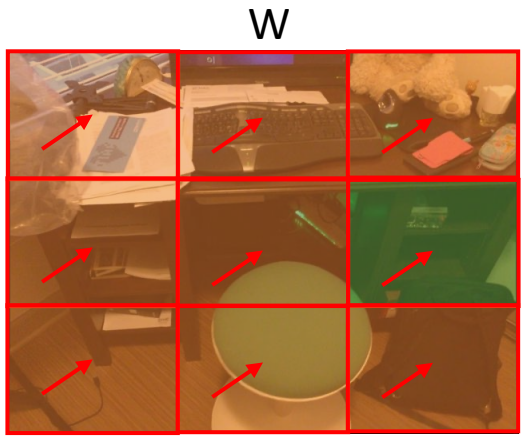


How to Perform Dense Matching?



How to Pretrain the Image Transformer?





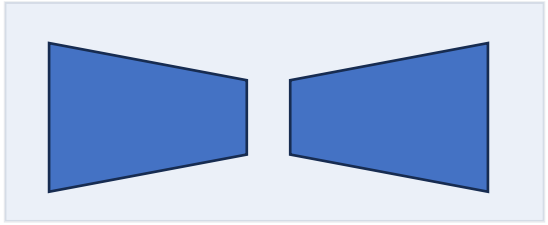


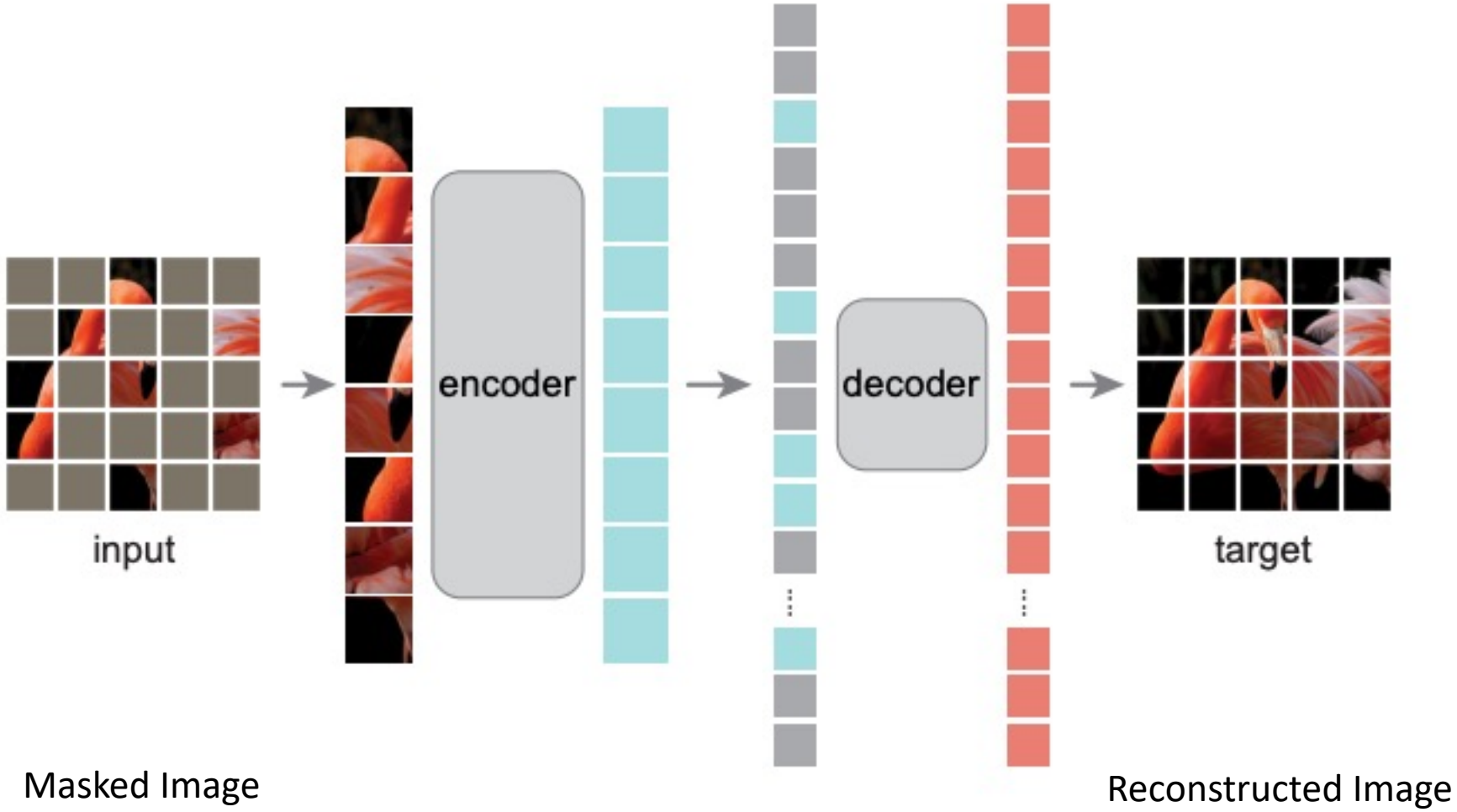
Image Transformer

Monocular Classification

Can Not Pretrain

ImageNet Classification

Background: Masked Image Modeling

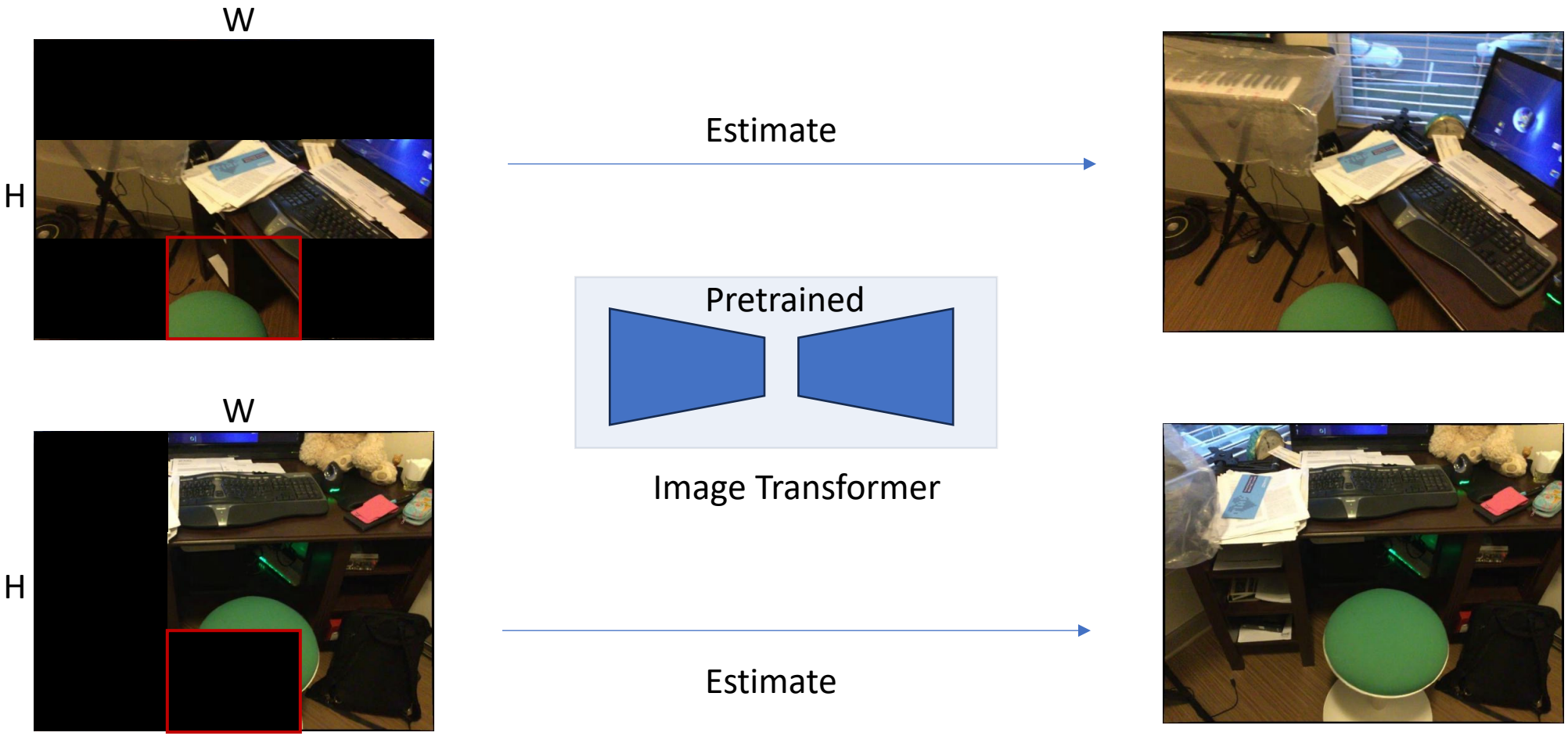


Masked Image

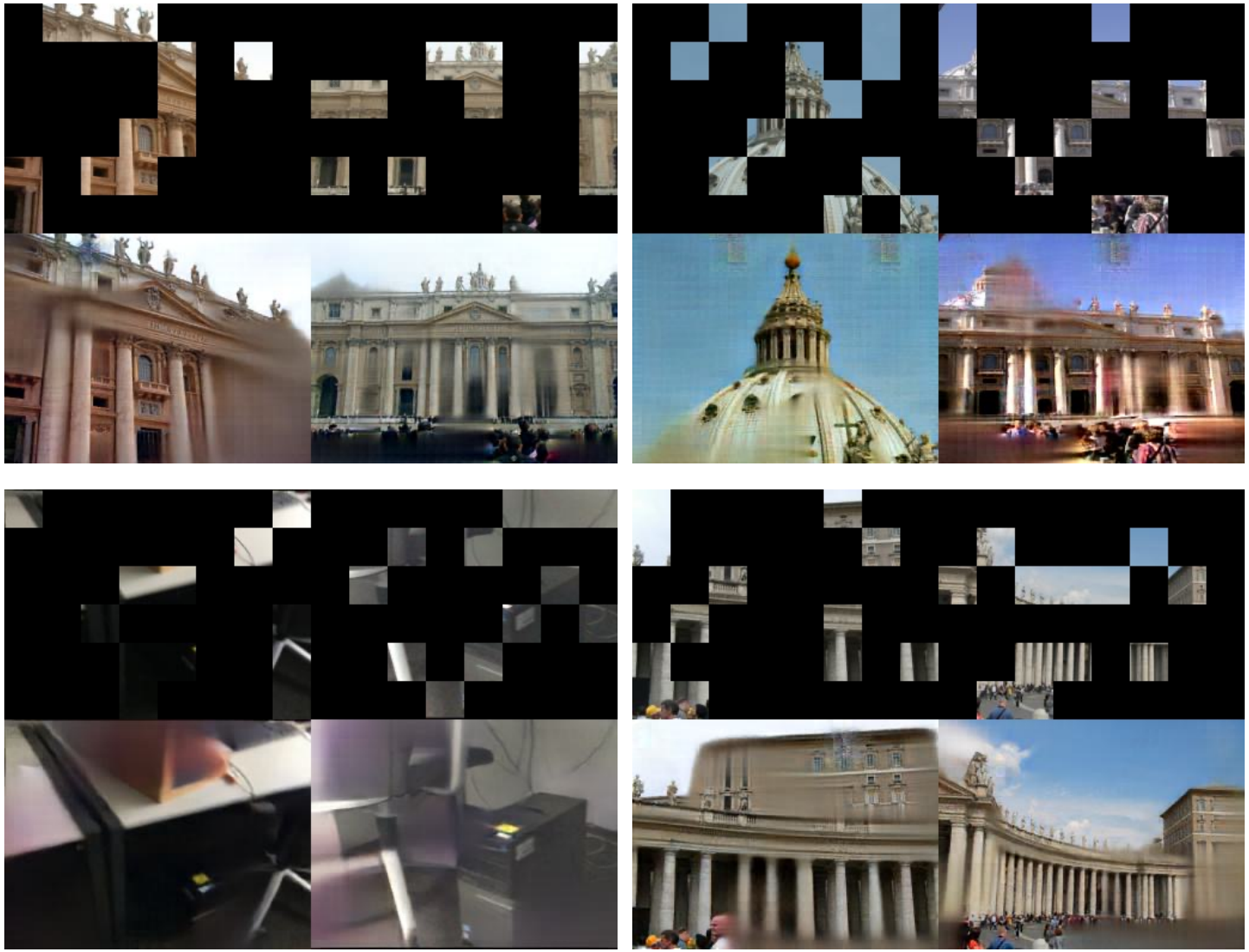
Reconstructed Image

Image Courtesy: https://arxiv.org/pdf/2111.06377.pdf

Masked Image Modeling to Paired Masked Image Modeling

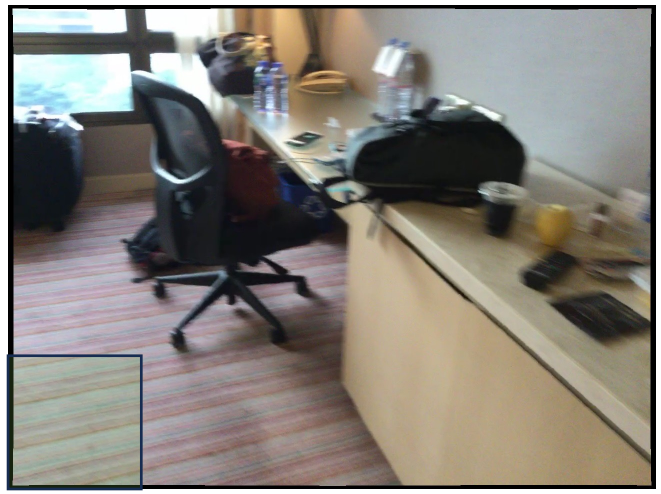


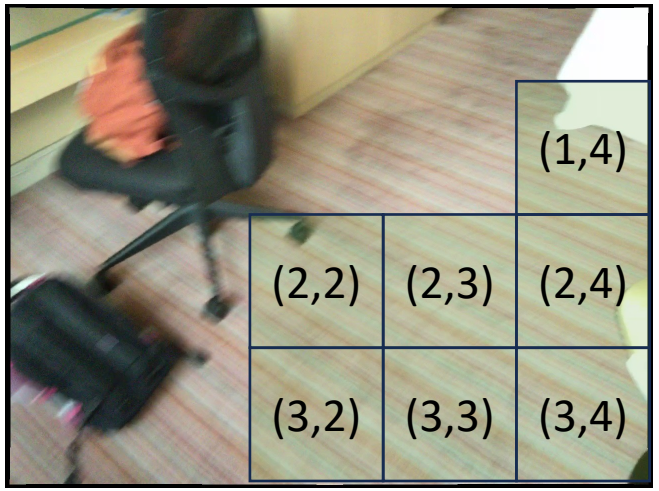
PMIM Pretexting Visual Quality



Address SoftMax Matching Ambiguity

Ambiguity in SoftMax Exhaustive Matching



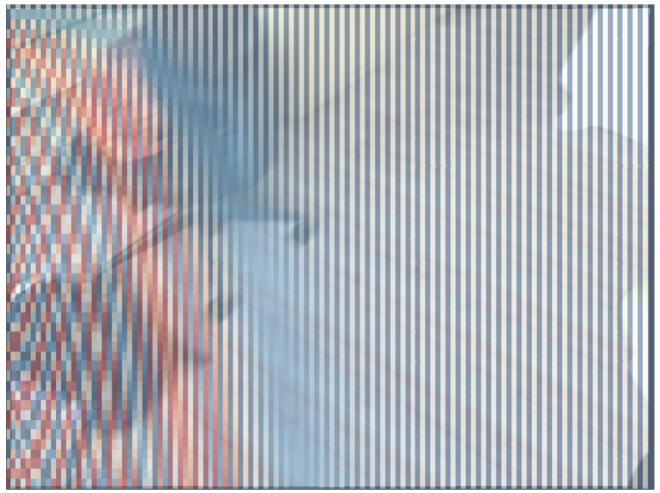


- SoftMaxed Matching Encounters Ambiguity when Multiple Similar Patches Present

$$\mathbf{x} = \frac{1}{7}((1,4) + (2,2) + (2,3) + (2,4) + (3,2) + (3,3) + (3,4))$$

- How to Address the Ambiguity?

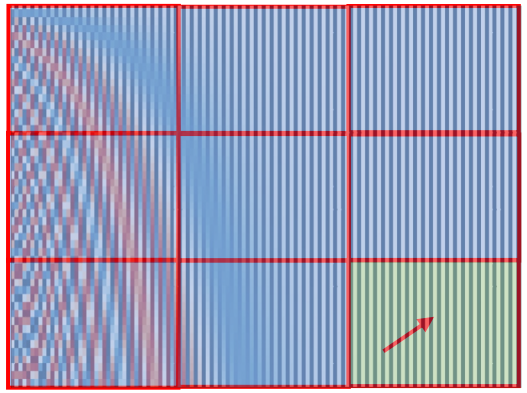
Address Ambiguity by Positional Embedding



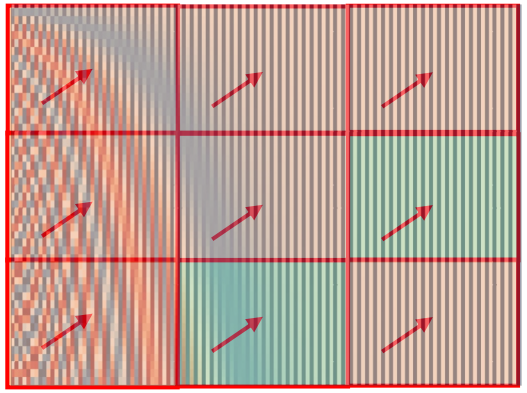
- Impose Positional Embedding
- Conduct Matching with Positional Embedding

Addrees Ambiguity by Positional Embedding

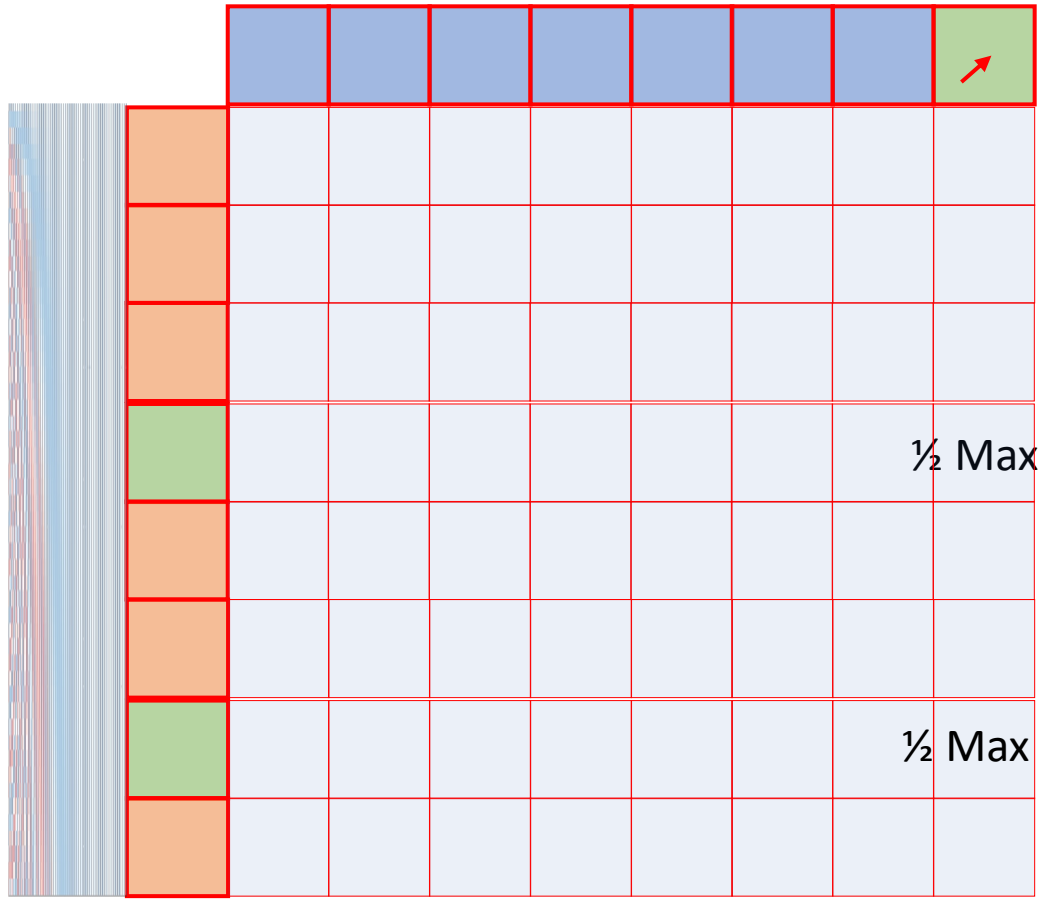
Feature Embedding 1



Feature Embedding 2



Exhaustive Matching



Source

$(3, 3)$

Original SoftMax →

Target

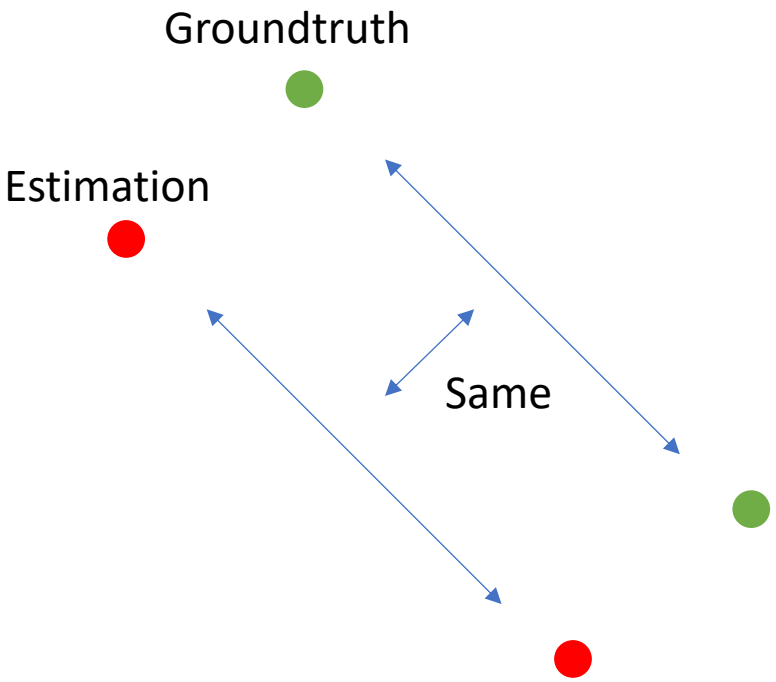
$$\frac{1}{2}((2,3)+(3,2))$$

With Embedding →

$$\frac{1}{2}\left(\mathbf{v}_2^3+\mathbf{v}_3^2\right)$$

Decoder →

Address Ambiguity by Planar Constraint



- Planar Structure Leads to Low-DoF Homography Transformation
- Learn Planar Prior with First-Order Loss
- Distance Between Pairs are same

Result

Performance

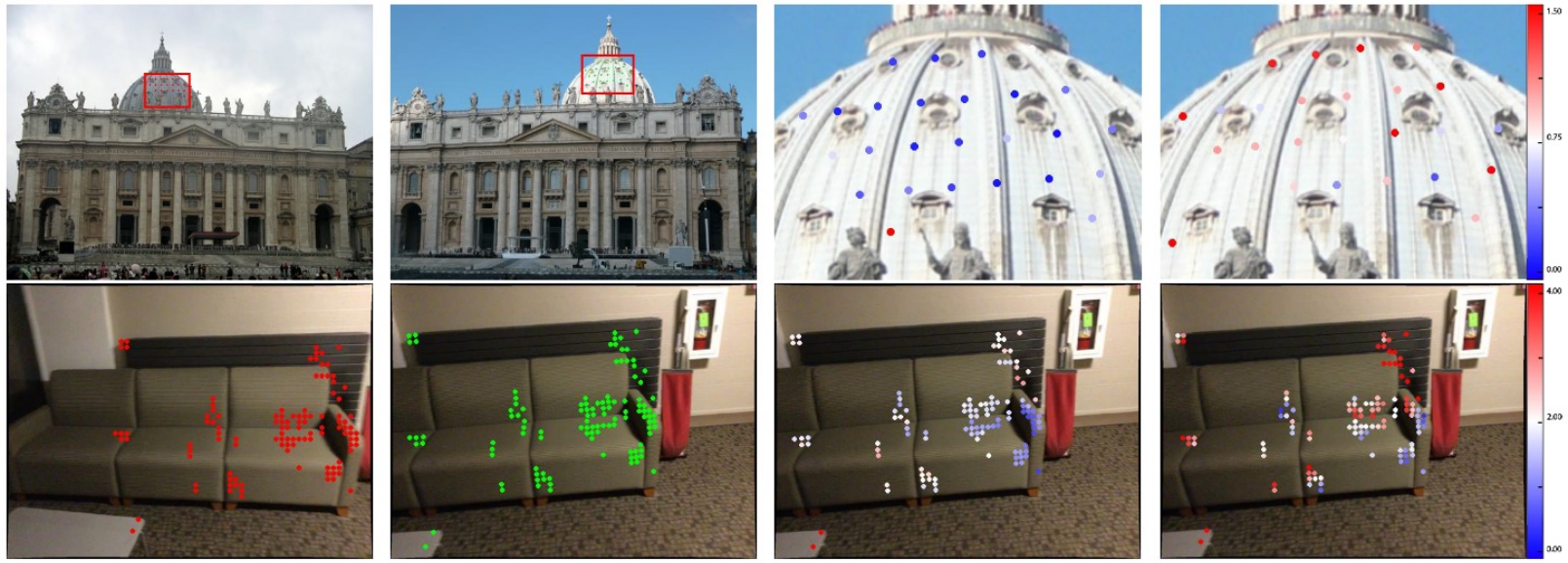
Category	Methods	Venue	Pose Estimation AUC ↑		
			@5°	@10°	@20°
Sparse W/ Detector	SuperGlue	CVPR'19	42.2	61.2	75.9
	SGMNet	Pattern'20	40.5	59.0	72.6
Sparse Wo/ Detector	DRC-Net	ICASSP'22	27.0	42.9	58.3
	LoFTR	CVPR'21	52.8	69.2	81.2
	QuadTree	ICLR'22	54.6	70.5	82.2
	MatchFormer	ACCV'22	53.3	69.7	81.8
	ASpanFormer	ECCV'22	**55.3**	**71.5**	**83.1**
Dense	PDC-Net+	Arxiv'19	43.1	61.9	76.1
	PMatch (Ours)	CVPR'23	**61.4**	**75.7**	**85.7**

- MegaDepth Performance

Category	Methods	Venue	Pose Estimation AUC ↑		
			@5°	@10°	@20°
Sparse W/ Detector	SuperGlue	CVPR'19	16.2	33.8	51.8
	SGMNet	PR'20	15.4	32.1	48.3
Sparse Wo/ Detector	DRC-Net	ICASSP'22	7.7	17.9	30.5
	LoFTR	CVPR'21	22.0	40.8	57.6
	QuadTree	ICLR'22	24.9	44.7	61.8
	MatchFormer	ACCV'22	24.3	43.9	61.4
	ASpanFormer	ECCV'22	**25.6**	**46.0**	**63.3**
Dense	PDC-Net+	Arxiv'19	20.2	39.4	57.1
	PMatch (Ours)	CVPR'23	**29.4**	**50.1**	**67.4**

- ScanNet Performance

PMIM Visual Quality



Source Support LoFTR Ours

Thanks For Watching!