JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

# Structure Aggregation for Cross-Spectral Stereo Image Guided Denoising

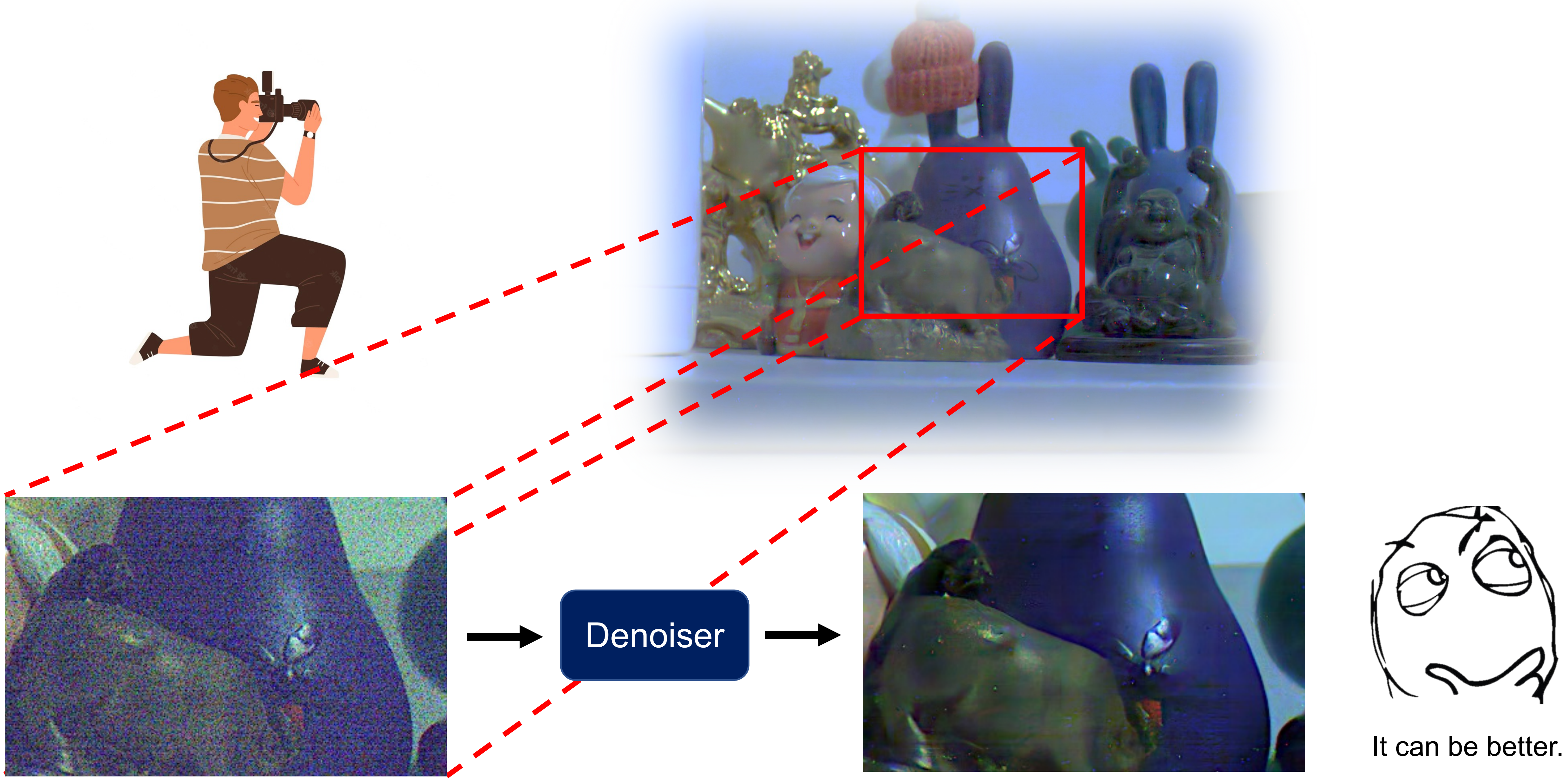

Zehua Sheng[1], Zhu Yu[1], Xiongwei Liu[1], Si-Yuan Cao[1], Yuqi Liu[1], Hui-Liang Shen[1], Huaqi Zhang[2]

[1]Zhejiang University

[2]vivo Communication Company Ltd.


Poster: WED-PM-157

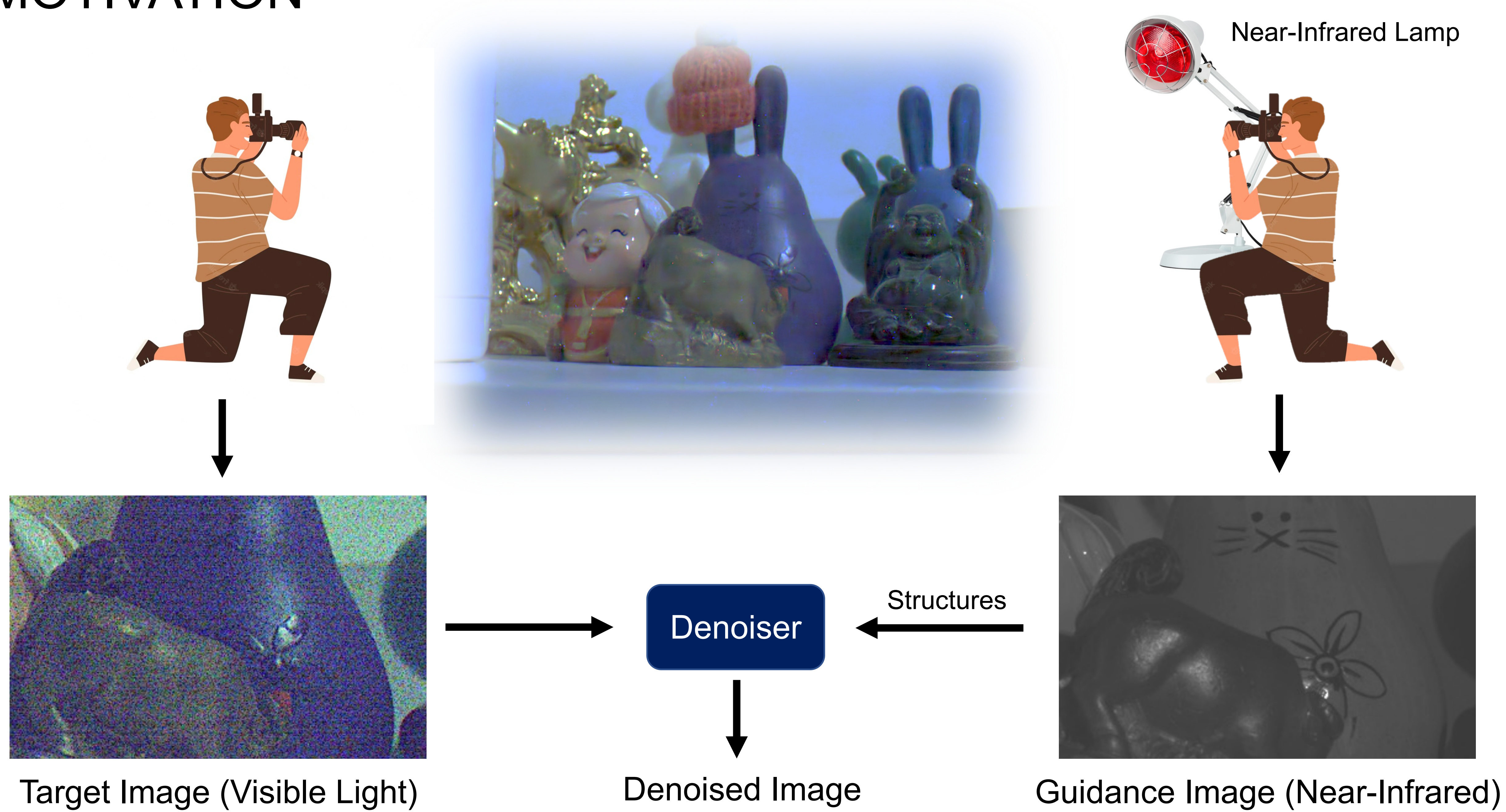
# MOTIVATION



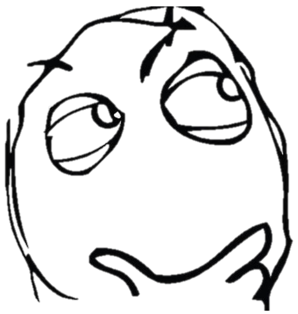

Noisy Image

Denoised Image

# MOTIVATION

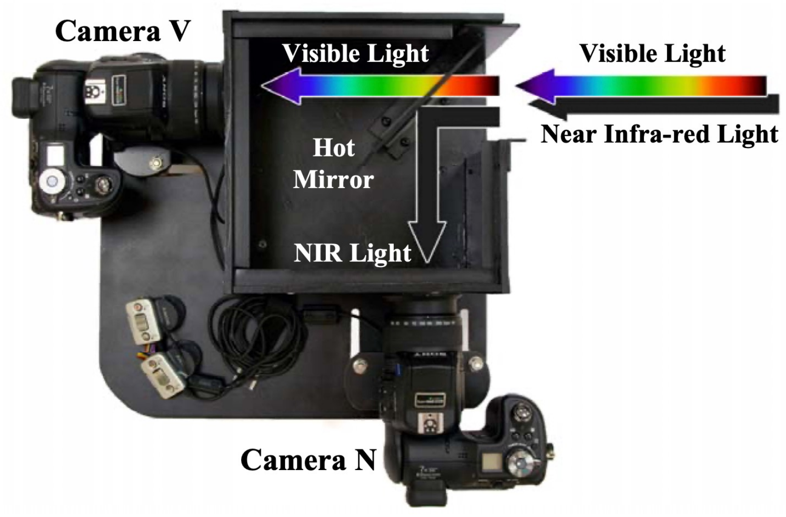

Near-Infrared Lamp
Denoiser
Structures
Target Image (Visible Light)
Denoised Image
Guidance Image (Near-Infrared)

# MOTIVATION

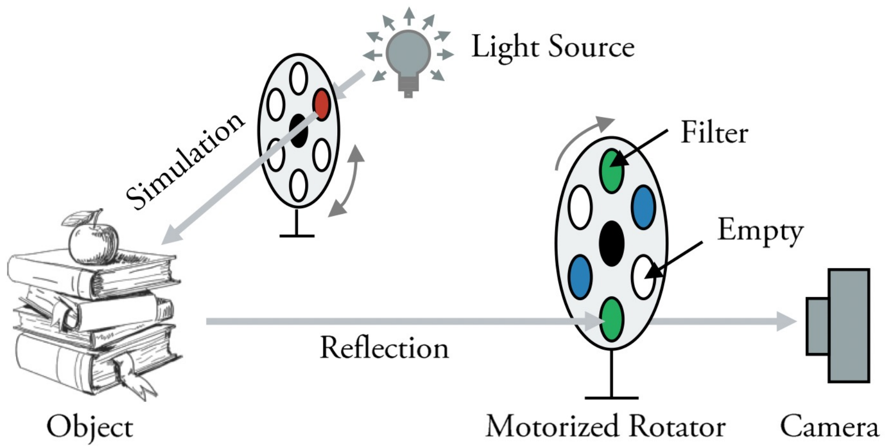


How do we obtain paired target and guidance images?

But how can we apply them into portable devices?




with Beam Splitter [1]



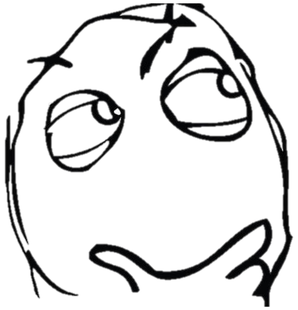

with Rotator [2]

Portable Devices

[1] Enhancing photographs with near infra-red images. -CVPR'08

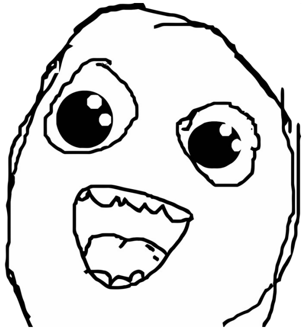
[2] An integrated enhancement solution for 24-hour colorful imaging. -AAAI'20

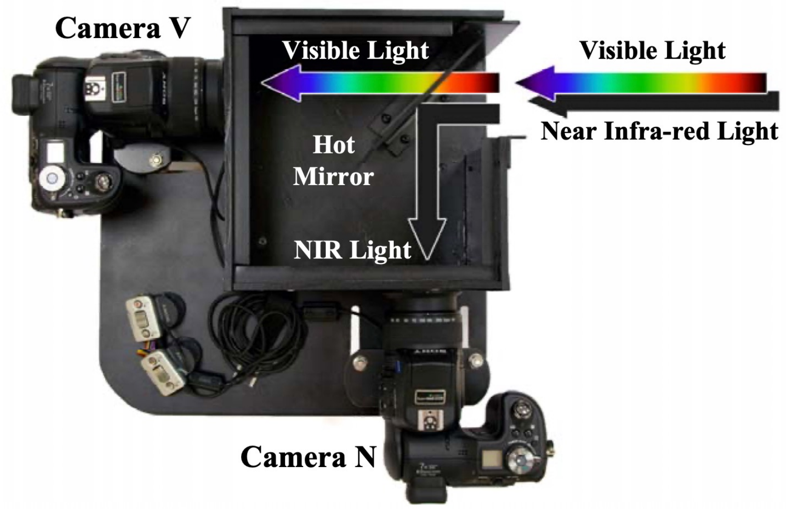
# MOTIVATION



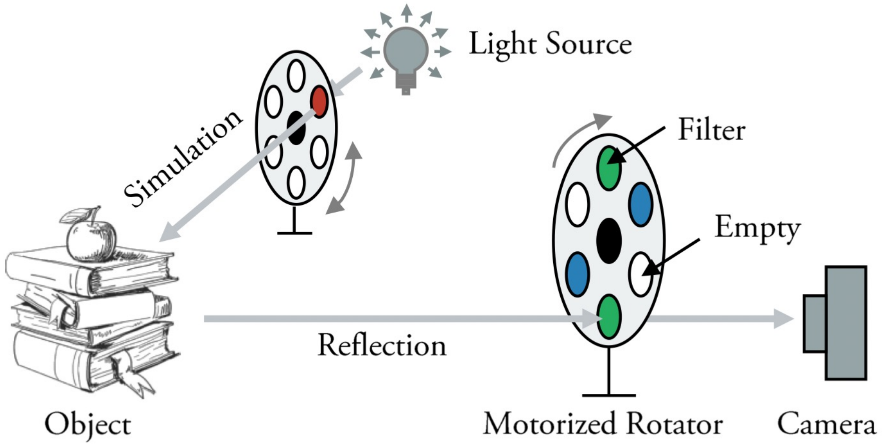
How do we obtain paired target and guidance images?

Stereo system.



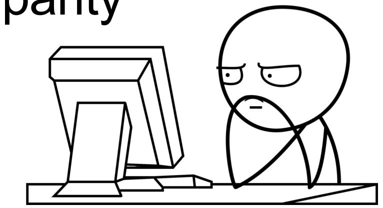



with Beam Splitter [1]




with Rotator [2]

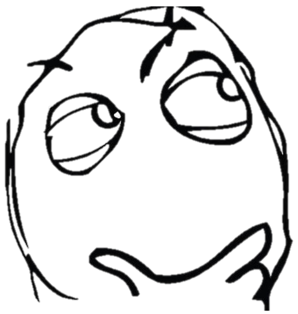
Disparity



[1] Enhancing photographs with near infra-red images. -CVPR'08

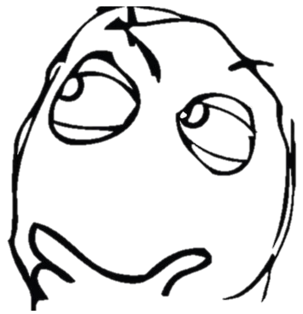
[2] An integrated enhancement solution for 24-hour colorful imaging. -AAAI'20

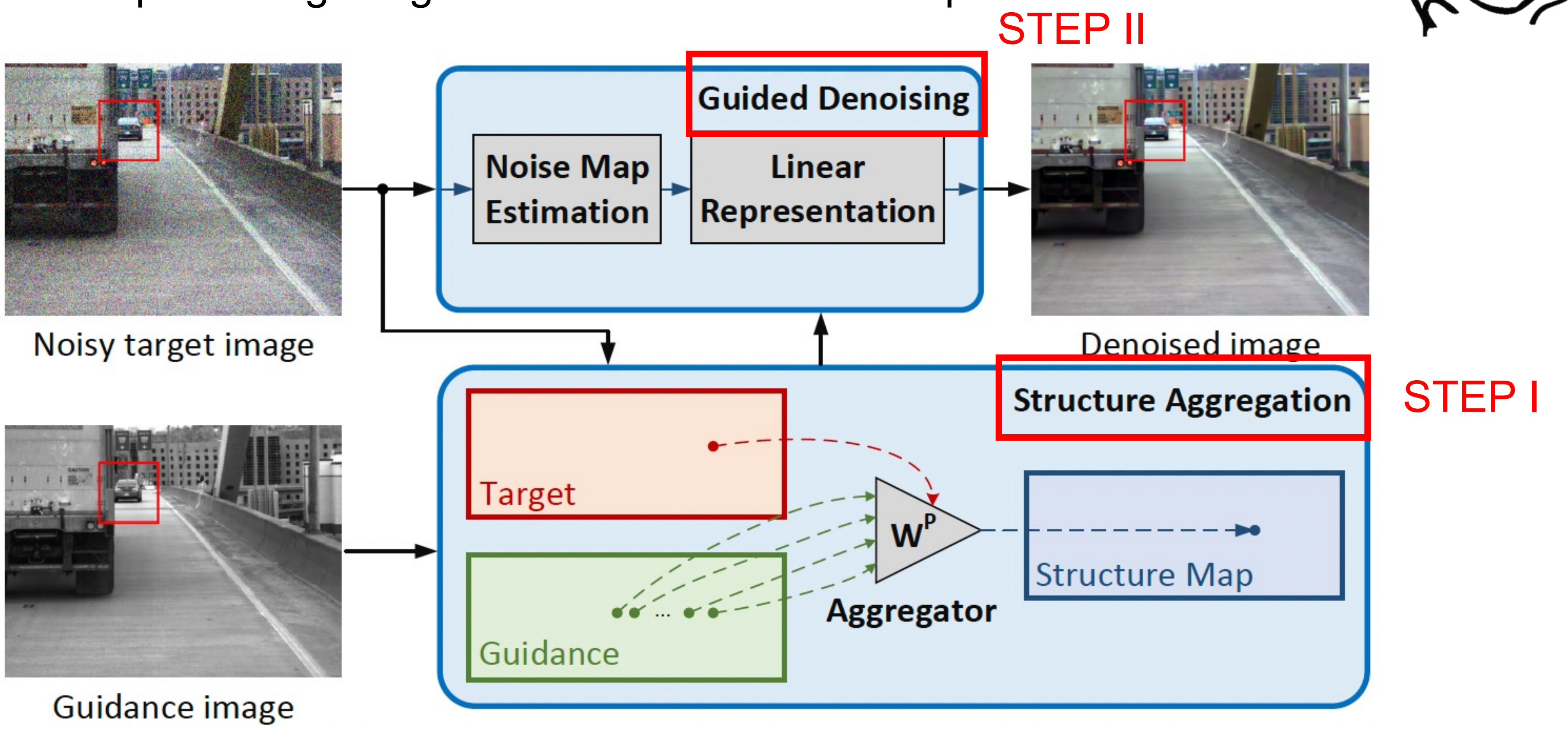
# OVERVIEW



How can we take the advantage of unaligned guidance information?

Considering the problem of cross-modality and image degradation, is explicit image alignment still an essential step?






Our proposed SANet

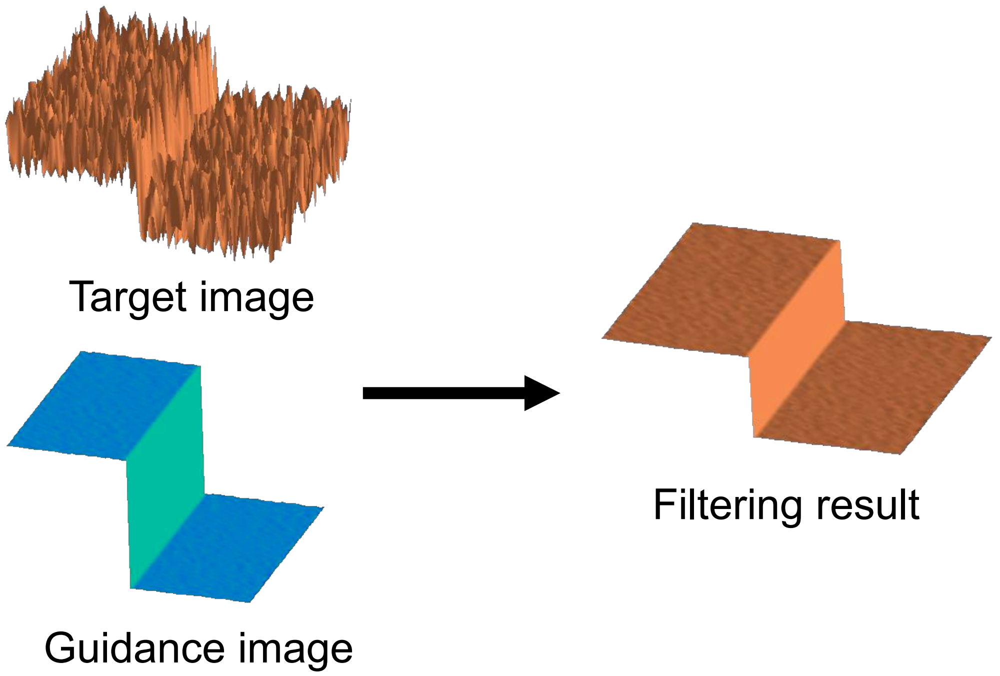
# REVISITING GUIDED FILTERING [1]

Given a noisy image $\mathbf{Y}$ and an aligned guidance image $\mathbf{G}$, the clean image $\mathbf{X}$ can be estimated by a spatially variant linear representation model of $\mathbf{G}$.

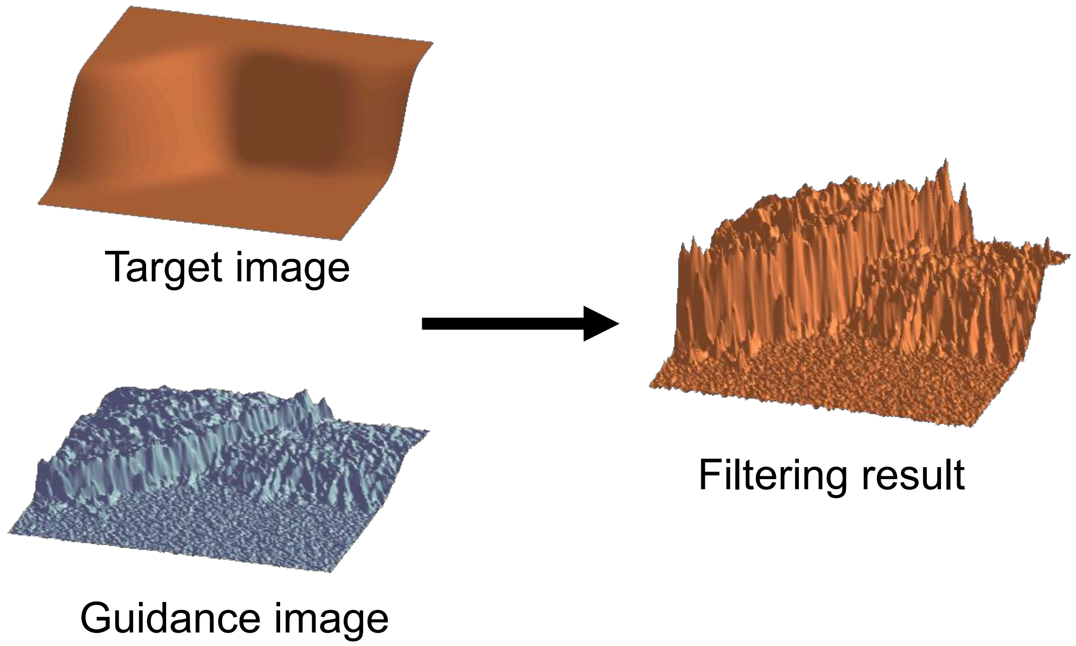
$$\widehat{\mathbf{X}} = \mathbf{A} \odot \mathbf{G} + \mathbf{B}$$

$\odot$: Hadamard Product

$$\widehat{\mathbf{X}}(i,j) = \mathbf{A}(i,j) \cdot \mathbf{G}(i,j) + \mathbf{B}(i,j)$$




Smoothing



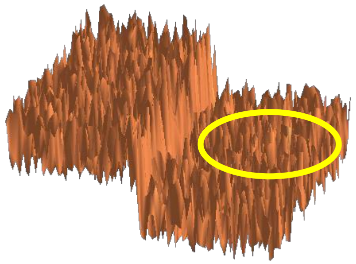

Detail Transferring

[1] Guided image filtering. -ECCV’10, TPMAI'12

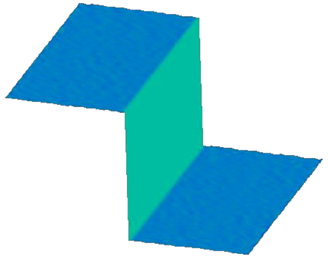
# REVISITING GUIDED FILTERING [1]

It's simple and effective. But it requires input images to be structurally consistent.

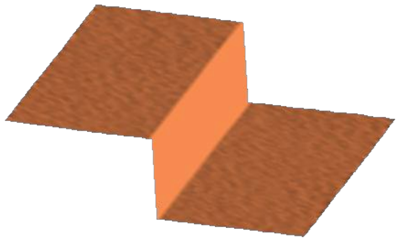
What if there are unique structures in the target image?



Target image

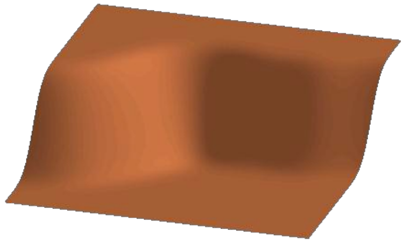


Guidance image

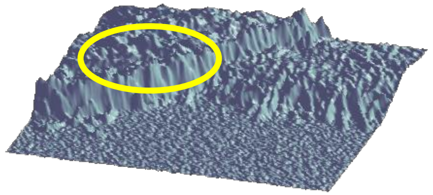


Filtering result

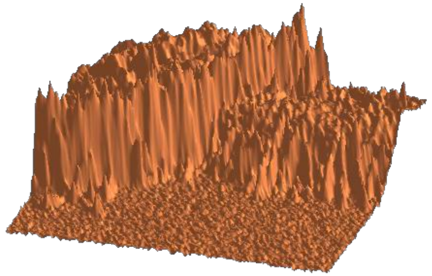
Smoothing



Target image



Guidance image



Filtering result

What if there are unique structures in the guidance image?

Detail Transfer

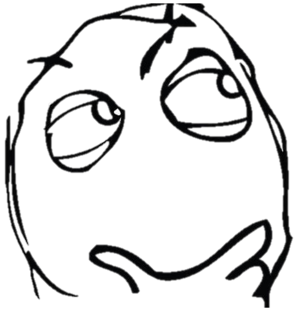
$$\hat{\mathbf{X}}(i,j) = \mathbf{A}(i,j) \cdot \mathbf{G}(i,j) + \mathbf{B}(i,j)$$

Weight $\mathbf{A}$ has the function of transferring structures from the guidance image, but struggles to judge whether the transfer is appropriate.

[1] Guided image filtering. -ECCV'10, TPMAI'12

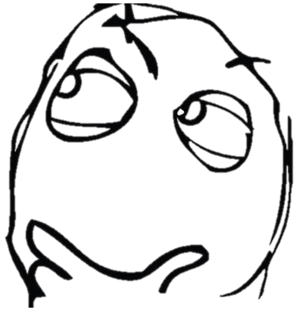
# PROBLEM FORMULATION

For stereo camera systems, the captured image pairs are structurally inconsistent due to disparity.

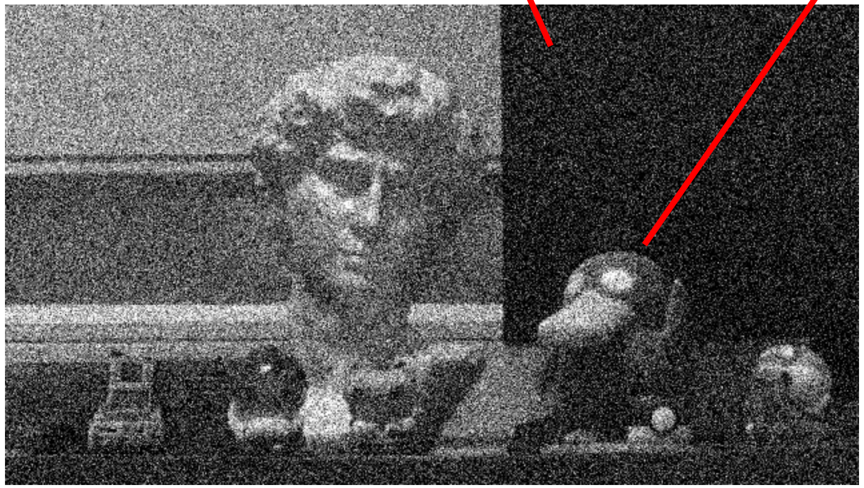


How do we obtain structurally aligned image pairs?

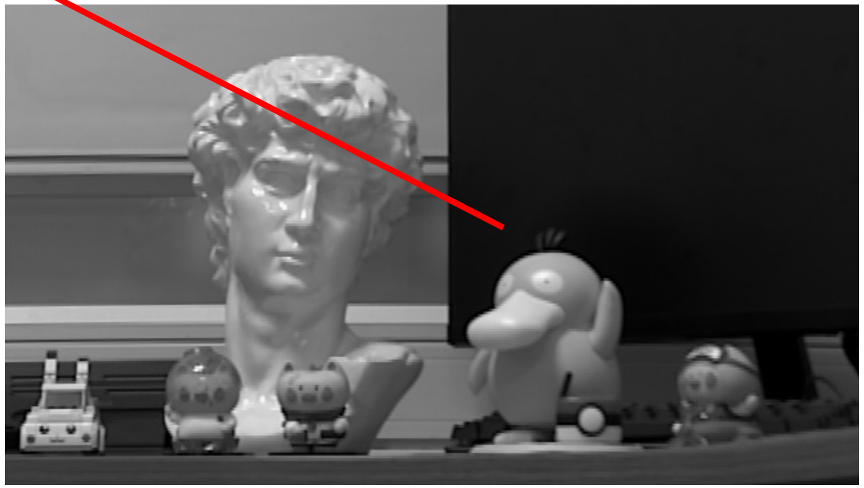
Stereo Matching?



Challenges: Degradation Cross-Modal/Spectral Disparity Supervision

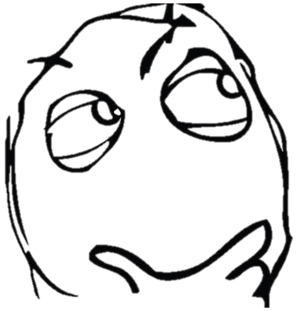


Target Image



Guidance Image

# PROBLEM FORMULATION



Our real purpose is to obtain two structurally aligned images.

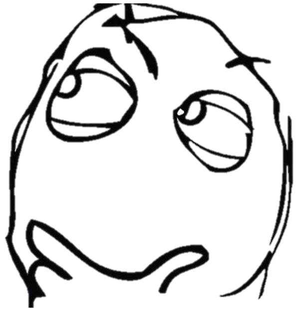
Do we still need to obey the one-to-one correspondence policy in stereo matching?


Target Image

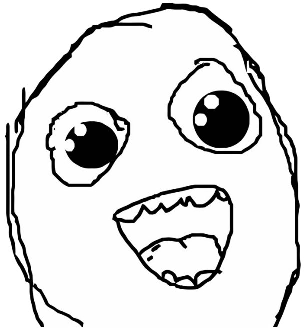

Guidance Image

The corresponding pixel in the guidance image is among the candidates restricted by the maximum disparity.

Some candidates are actually quite similar due to the structural redundancy of natural images.



# PROBLEM FORMULATION



How about aggregating these candidates together rather than selecting a single pixel?

Our proposed guided denoising model:

$$\widehat{\mathbf{X}}(i,j) = \sum_{d=0}^{D} \mathbf{W}_d(i,j) \cdot \mathbf{G}(i-d,j) + \mathbf{B}(i,j)$$

To enable the aggregation process to focus more on structural correspondence:

Scale Weight

Perceptual Weight *a.k.a.* Structure Aggregator

$$\widehat{\mathbf{X}}(i,j) = \mathbf{W}^S(i,j) \cdot \sum_{d=0}^{D} \mathbf{W}_d^P(i,j) \cdot \mathbf{G}(i-d,j) + \mathbf{B}(i,j)$$

$$= \mathbf{W}^S(i,j) \cdot \boldsymbol{U}(i,j) + \mathbf{B}(i,j)$$

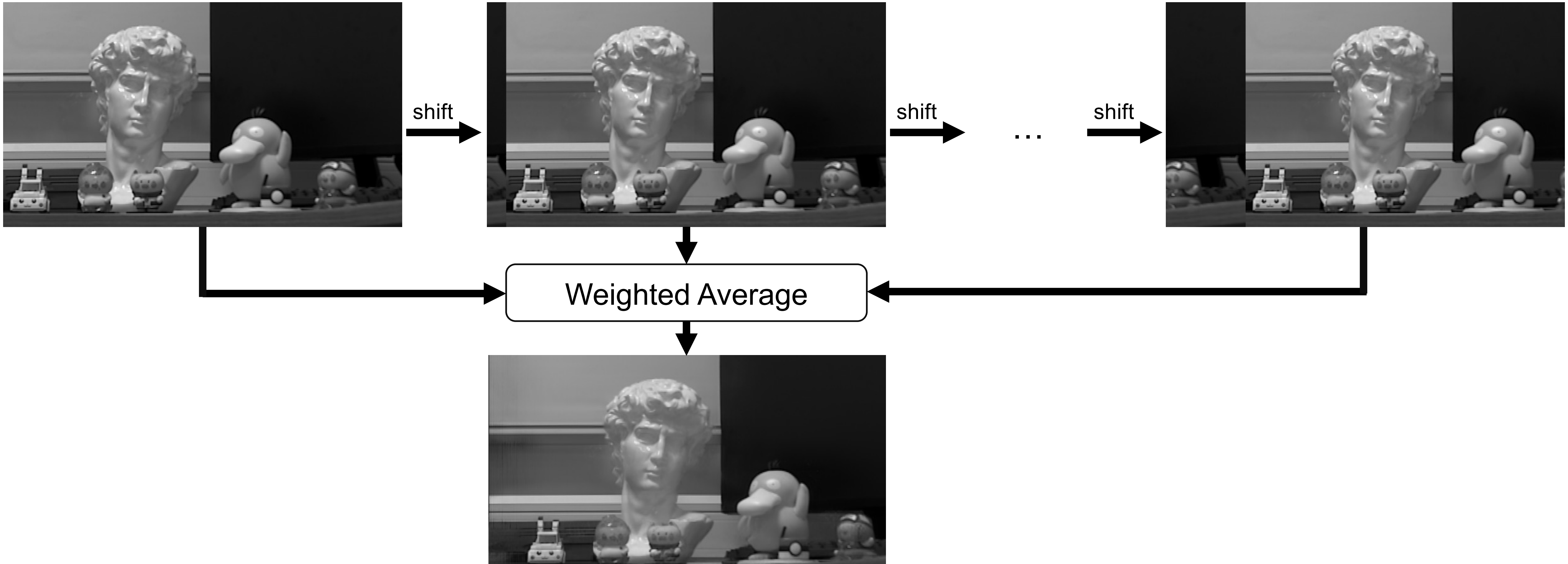
Structure Map

The denoising process is conducted in two stages: Structure Aggregation & Guided Denoising.

# STAGE I: STRUCTURE AGGREGATION

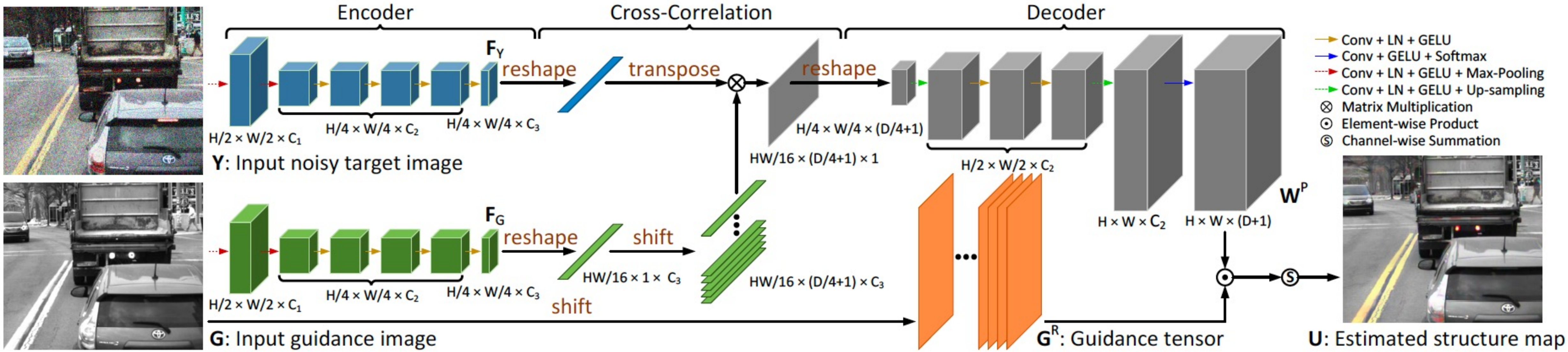
Aggregating non-local information to generate the structure map that is structurally aligned with the target image:

$$\mathbf{U}(i,j) = \sum_{d=0}^{D} \mathbf{W}_d^P(i,j) \cdot \mathbf{G}(i-d,j)$$



# STAGE I: STRUCTURE AGGREGATION

Network Architecture:

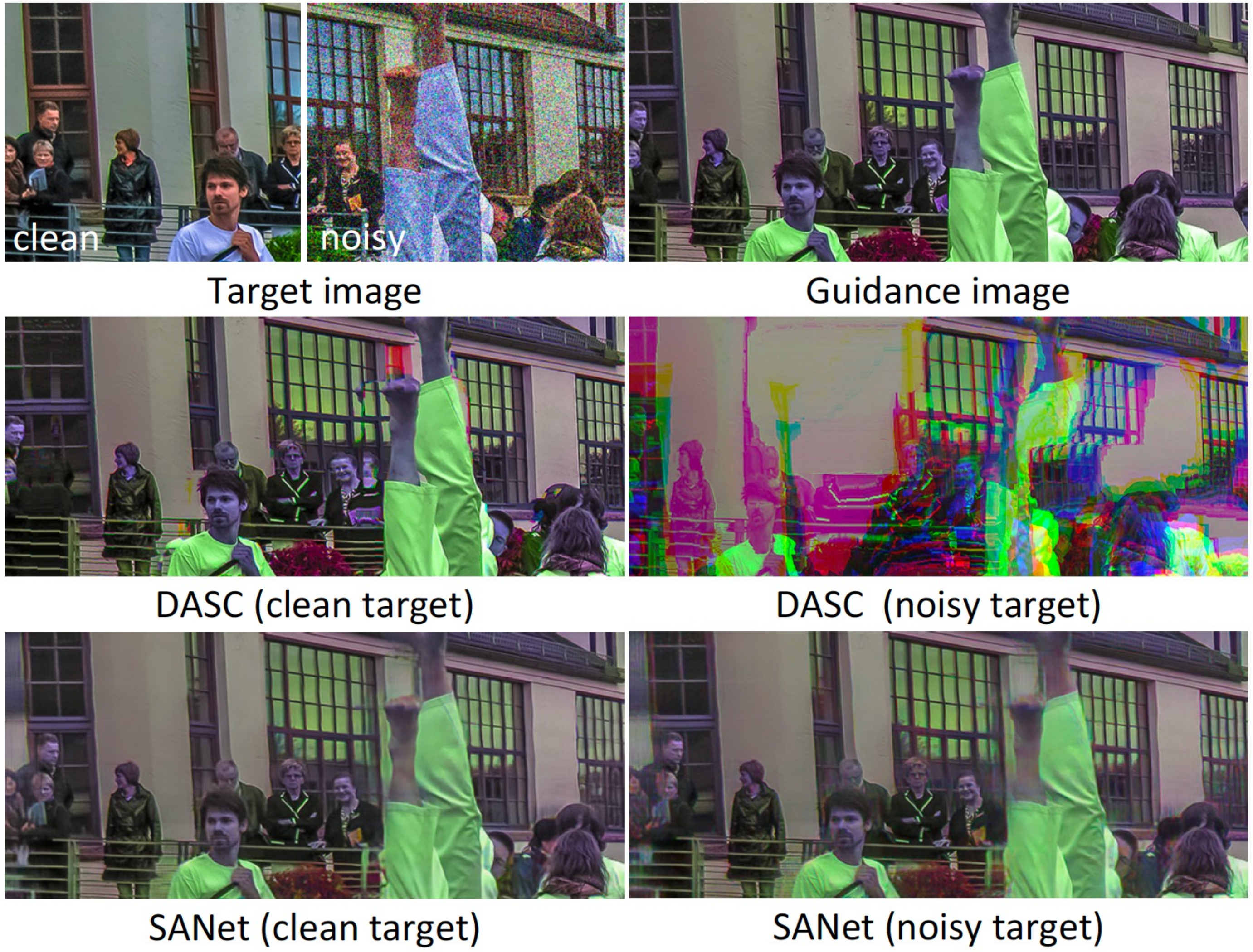



Loss Function:

$$\mathcal{L}_{SA}(\mathbf{U},\mathbf{X}) = \text{VGGPerceptual}(\mathbf{U},\mathbf{X})$$

# STAGE I: STRUCTURE AGGREGATION

Our structure aggregation strategy is robust to noise:




[DASC] DASC: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. -CVPR'15

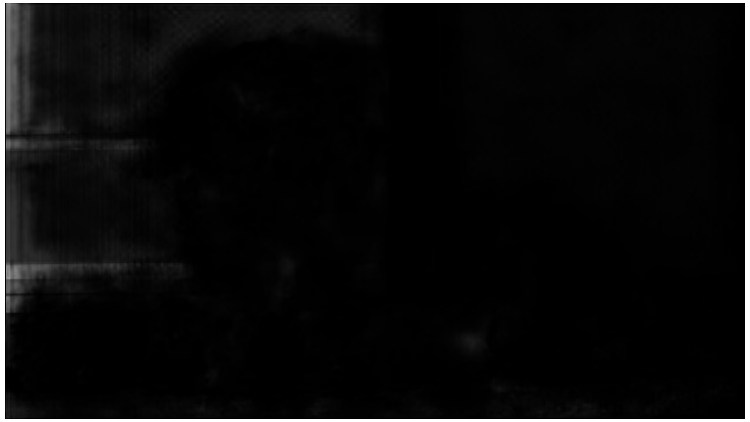
# AGGREGATION PROCESS


Shifted Guidance Image

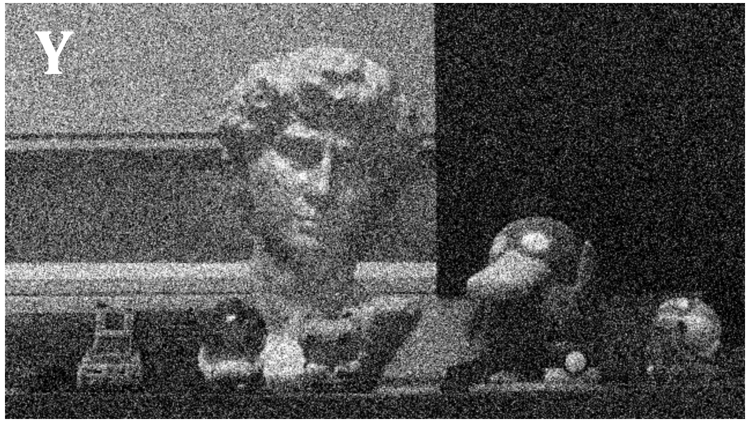
$\odot$


Perceptual Weight

$=$




$d = 0$




Noisy Target Image


Estimated Structure Map

# AGGREGATION PROCESS


Shifted Guidance Image

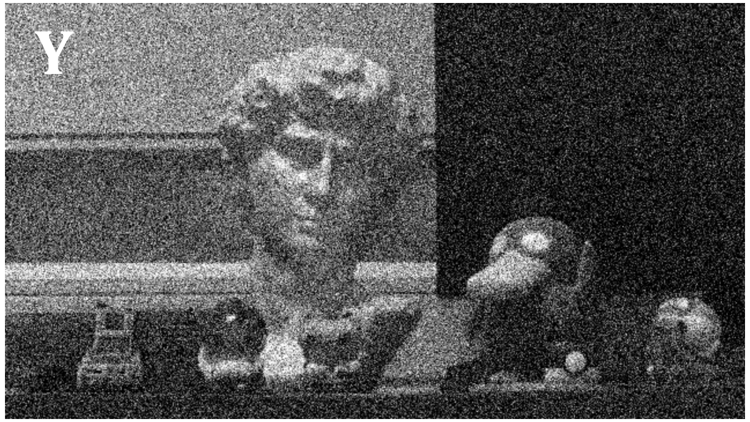
$\odot$


Perceptual Weight

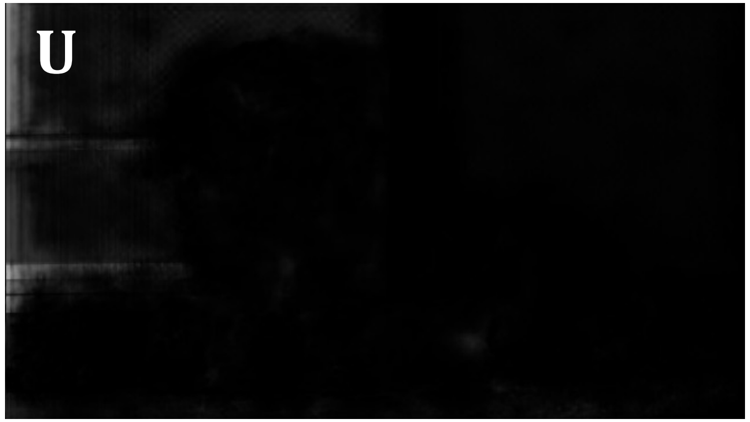
$=$

Aggregate

$d = 1$




Noisy Target Image




Estimated Structure Map

# AGGREGATION PROCESS


Shifted Guidance Image

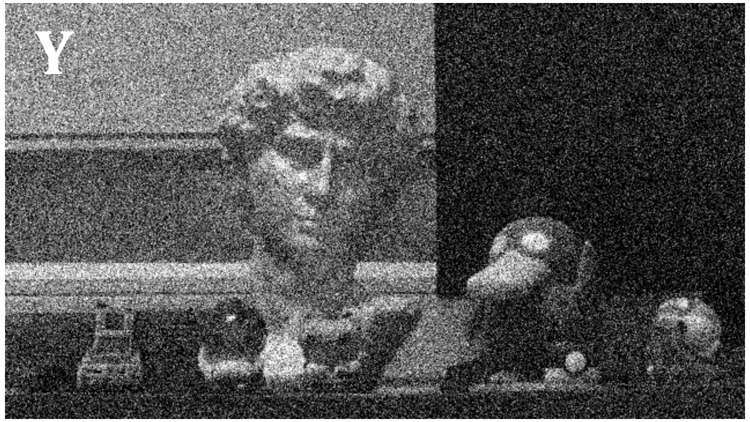
$\odot$


Perceptual Weight

$=$

Aggregate

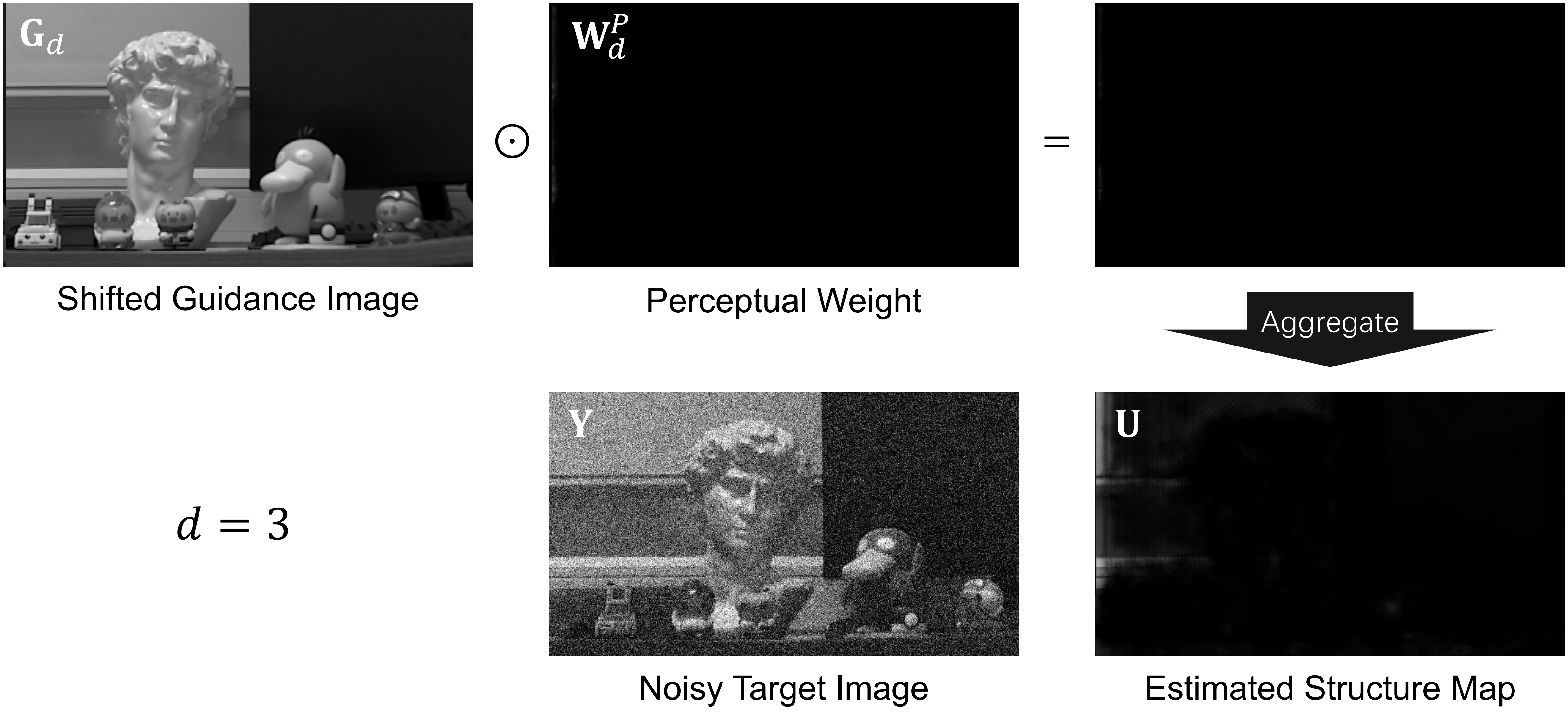
$d = 2$




Noisy Target Image


Estimated Structure Map

# AGGREGATION PROCESS



# AGGREGATION PROCESS


Shifted Guidance Image

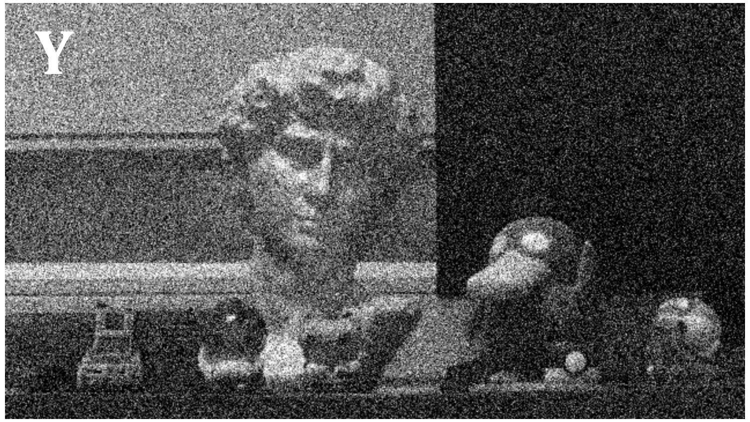
$\odot$


Perceptual Weight

$=$

Aggregate

$d = 4$




Noisy Target Image


Estimated Structure Map

# AGGREGATION PROCESS


Shifted Guidance Image

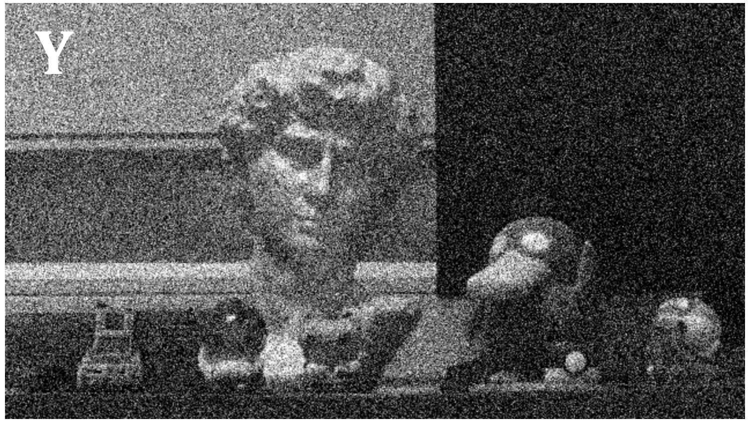
$\odot$


Perceptual Weight

$=$

Aggregate

$d = 5$




Noisy Target Image


Estimated Structure Map

# AGGREGATION PROCESS


Shifted Guidance Image

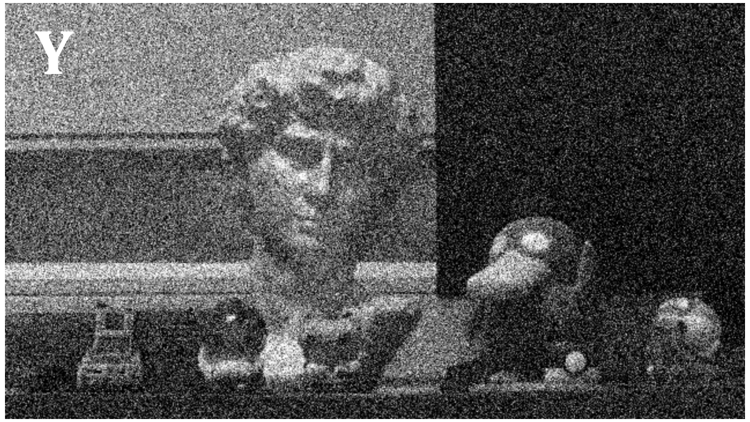
$\odot$


Perceptual Weight

$=$

Aggregate

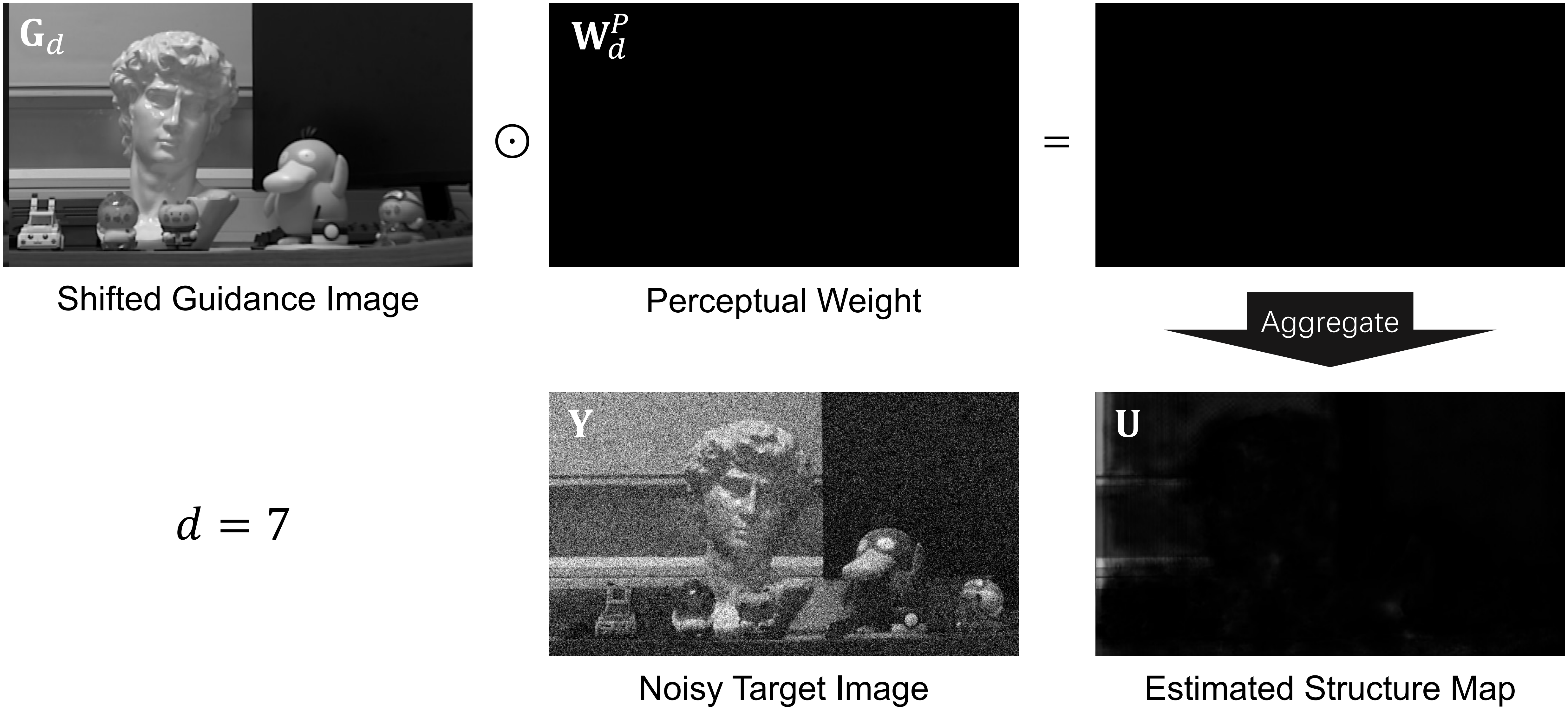
$d = 6$




Noisy Target Image


Estimated Structure Map

# AGGREGATION PROCESS



# AGGREGATION PROCESS


Shifted Guidance Image

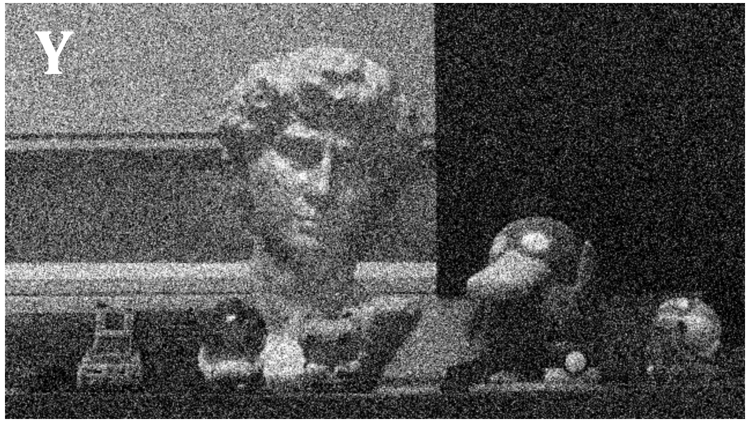
$\odot$


Perceptual Weight

$=$

Aggregate

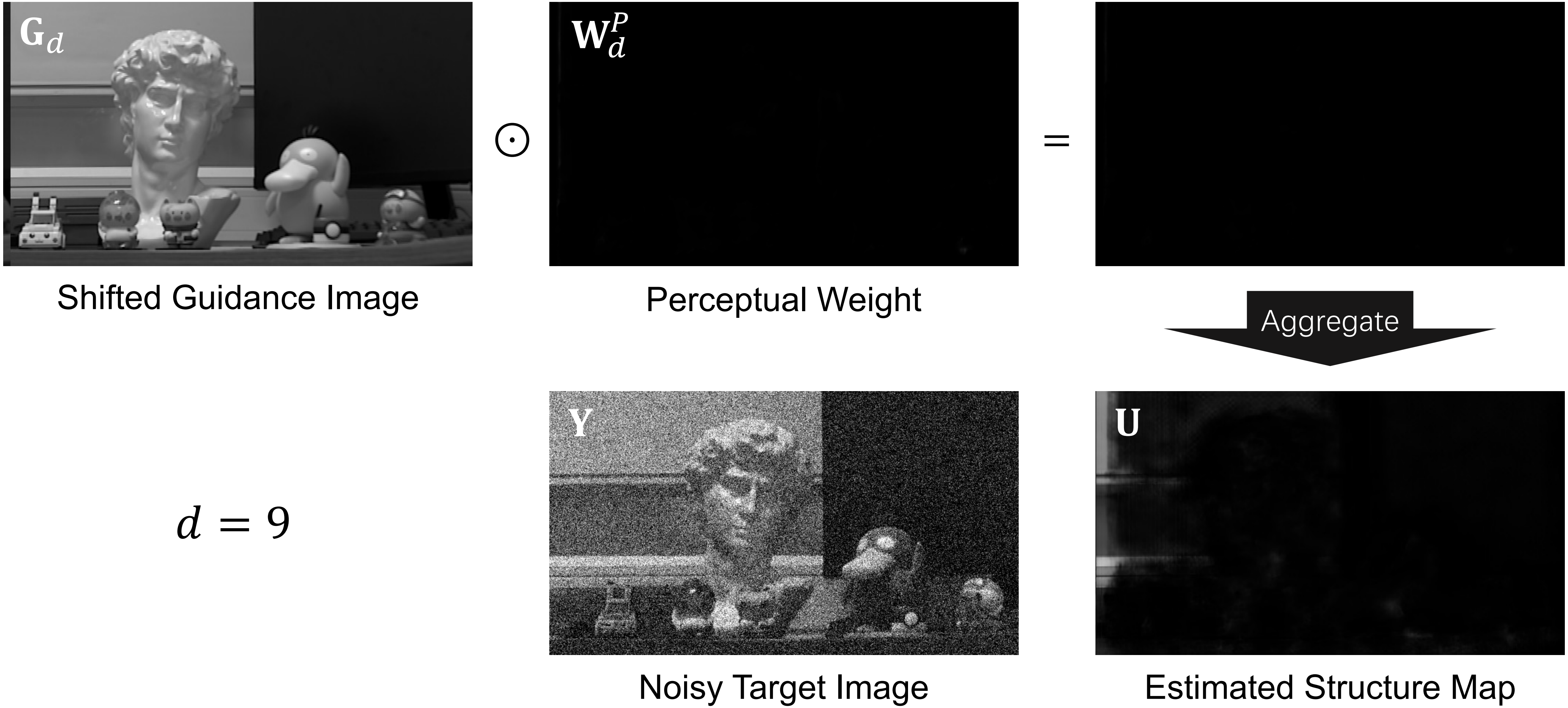
$d = 8$



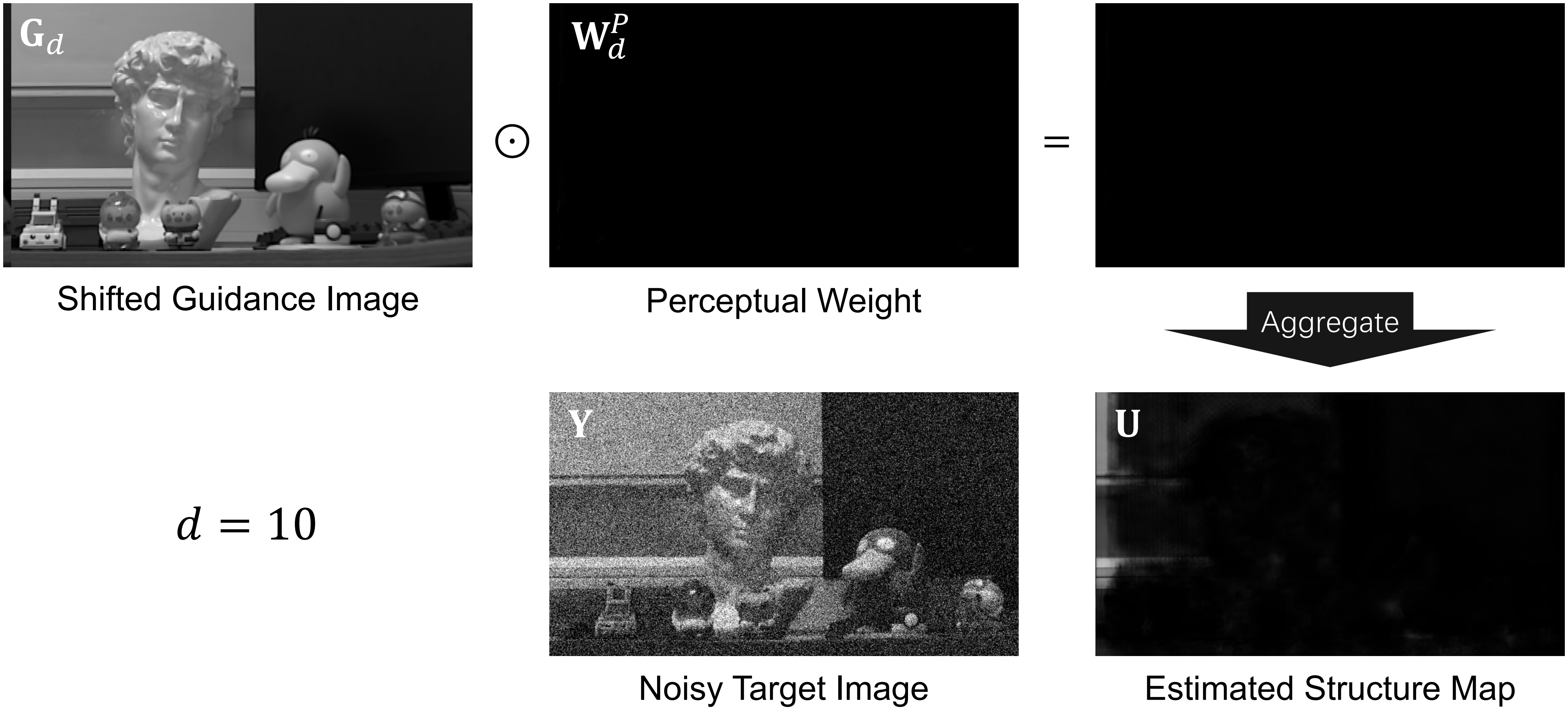

Noisy Target Image


Estimated Structure Map

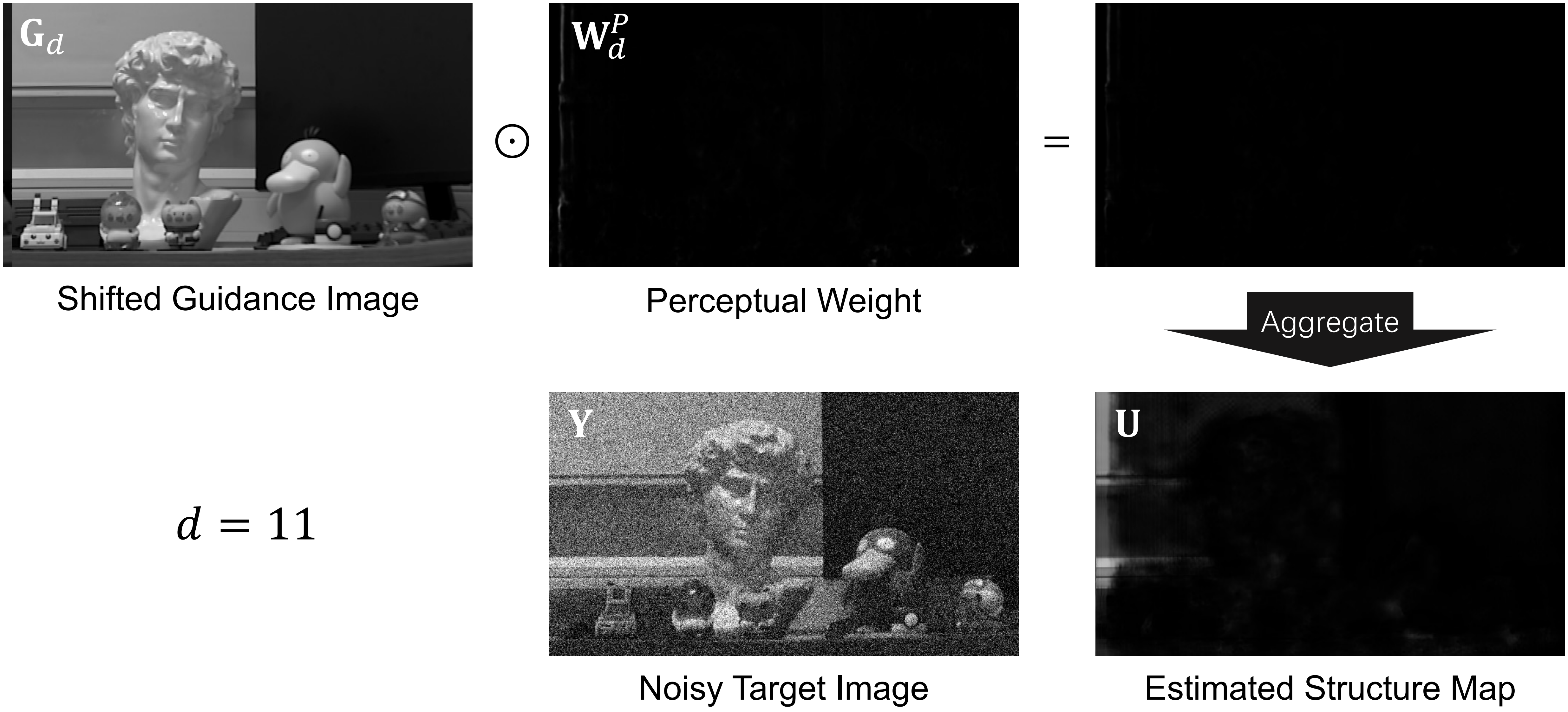
# AGGREGATION PROCESS



# AGGREGATION PROCESS



# AGGREGATION PROCESS



# AGGREGATION PROCESS


Shifted Guidance Image

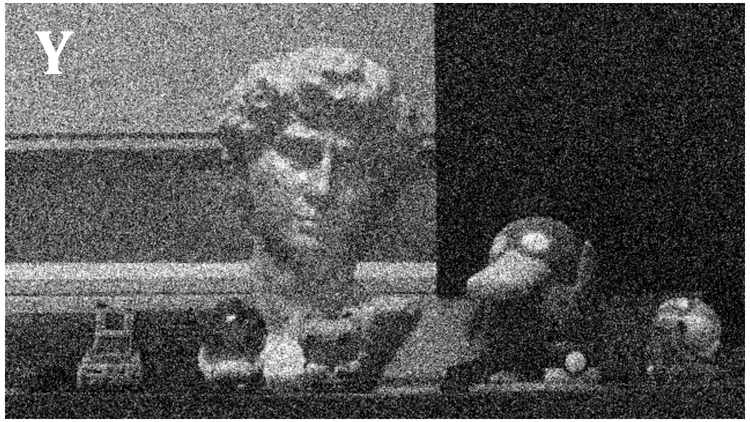
$\odot$


Perceptual Weight

$=$

Aggregate

$d = 12$




Noisy Target Image


Estimated Structure Map

# AGGREGATION PROCESS


Shifted Guidance Image

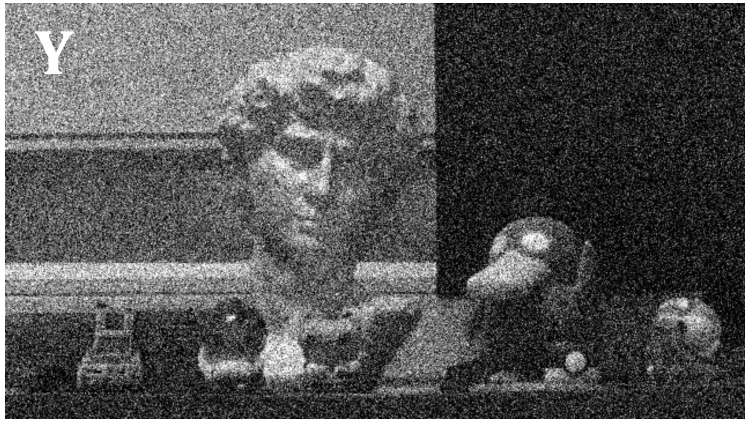
$\odot$


Perceptual Weight

$=$

Aggregate

$d = 13$




Noisy Target Image


Estimated Structure Map

# AGGREGATION PROCESS

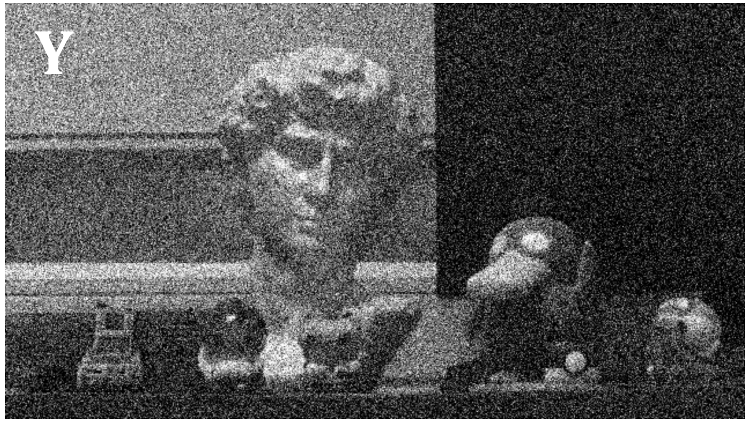

Shifted Guidance Image

$\odot$


Perceptual Weight

$=$


$d = 14$




Noisy Target Image


Estimated Structure Map

# AGGREGATION PROCESS


Shifted Guidance Image

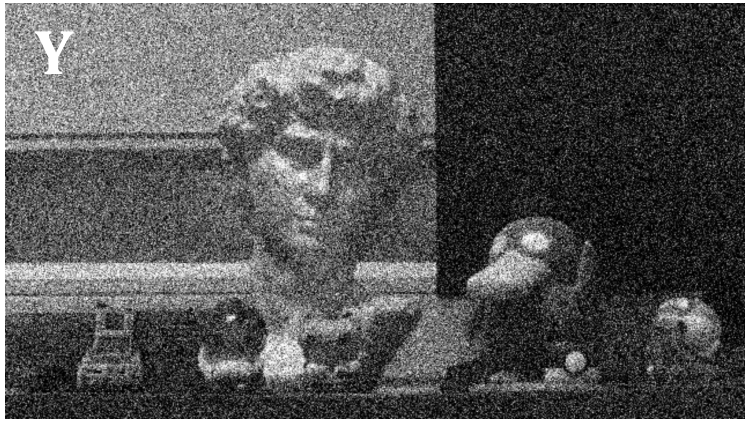
$\odot$


Perceptual Weight

$=$

Aggregate

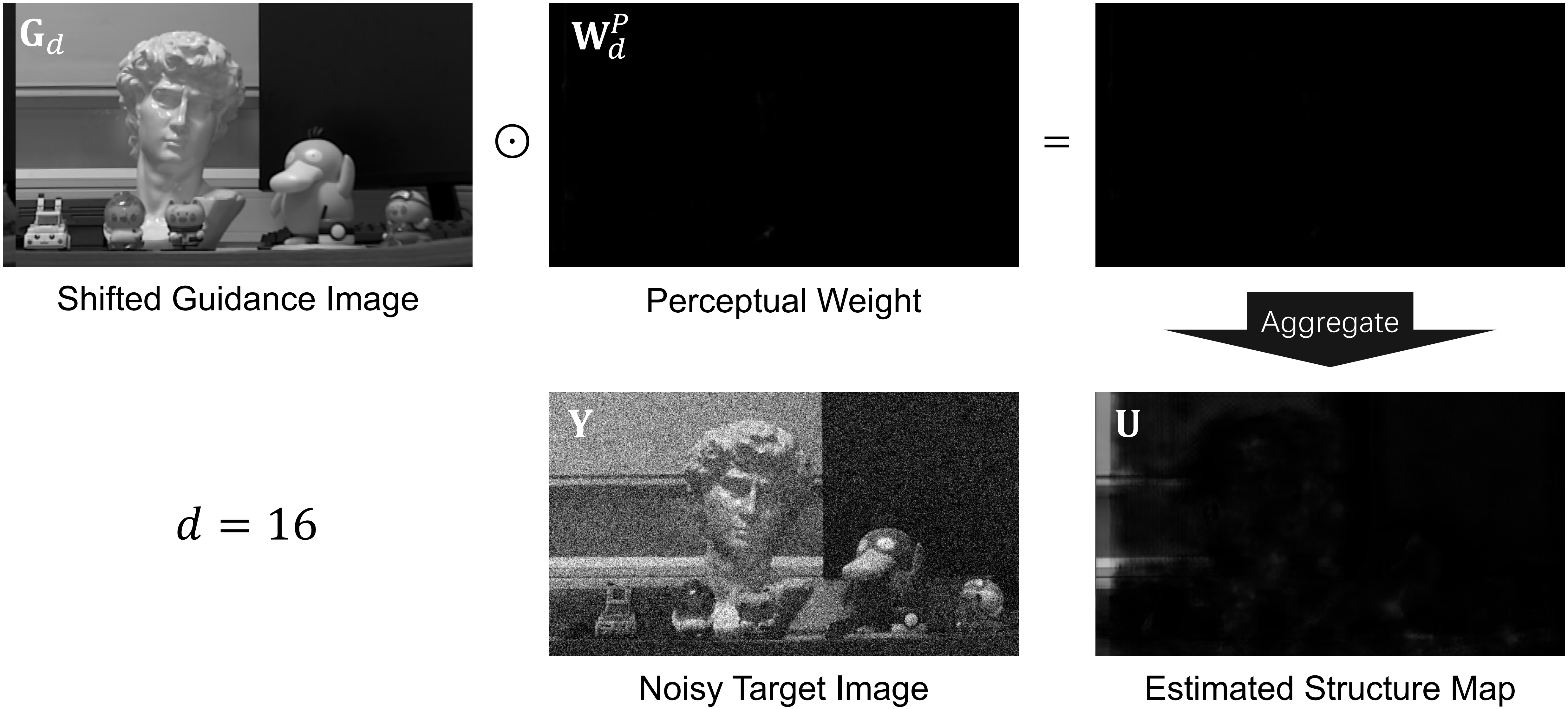
$d = 15$




Noisy Target Image


Estimated Structure Map

# AGGREGATION PROCESS



# AGGREGATION PROCESS


Shifted Guidance Image

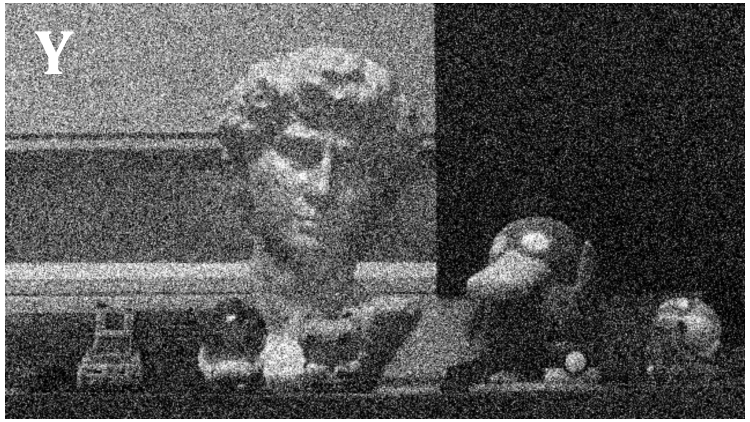
$\odot$


Perceptual Weight

$=$

Aggregate

$d = 17$




Noisy Target Image


Estimated Structure Map

# AGGREGATION PROCESS


Shifted Guidance Image

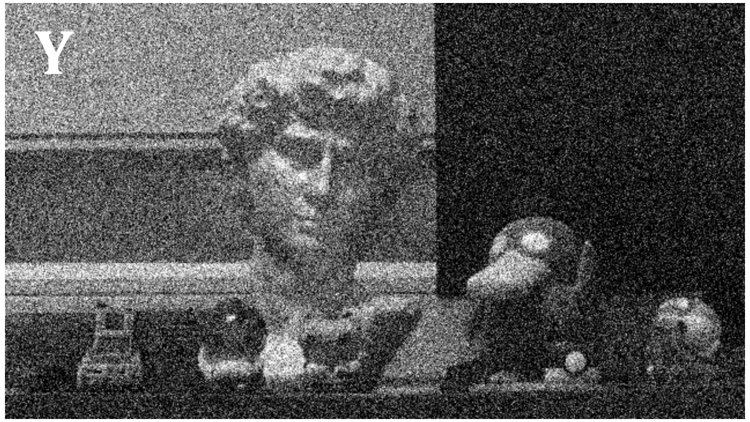
$\odot$


Perceptual Weight

$=$

Aggregate

$d = 18$




Noisy Target Image


Estimated Structure Map

# AGGREGATION PROCESS


Shifted Guidance Image

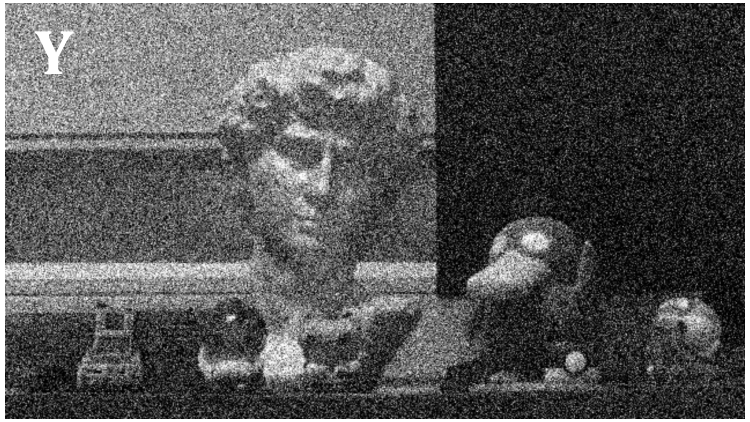
$\odot$


Perceptual Weight

$=$

Aggregate

$d = 19$




Noisy Target Image


Estimated Structure Map

# AGGREGATION PROCESS


Shifted Guidance Image

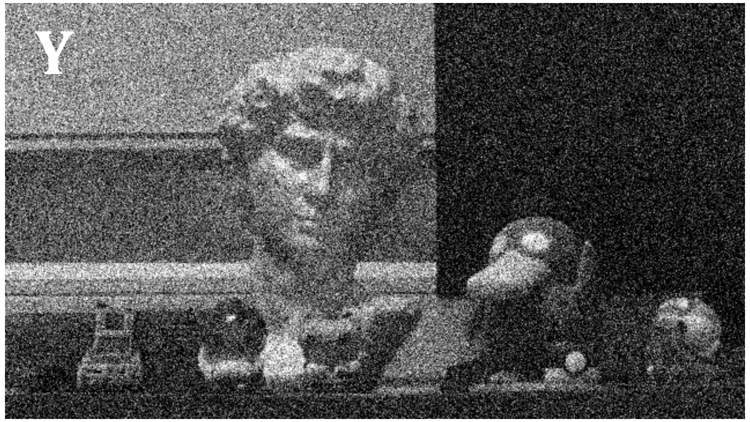
$\odot$


Perceptual Weight

$=$

Aggregate

$$d = 20$$




Noisy Target Image


Estimated Structure Map

# AGGREGATION PROCESS


Shifted Guidance Image

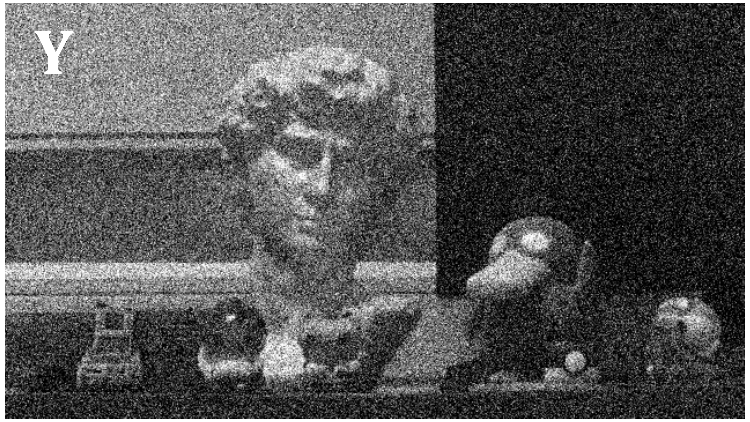
$\odot$


Perceptual Weight

$=$

Aggregate

$$d = 21$$




Noisy Target Image


Estimated Structure Map